REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
30-11-2005	Final Technical Report	24-03-2003 - 30-9-2005

4. TITLE AND SUBTITLE

Using Team Communication to Understand Team Cognition in Distributed vs. Co-located Mission Environments

5a. CONTRACT NUMBER

5b. GRANT NUMBER

N00014-03-1-0580

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Nancy J. Cooke, Jamie C. Gorman, Preston A. Kiekel, Peter Foltz, and Melanie Martin

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Arizona State University Polytechnic
7001 E. Williams Field Rd.
Mesa, AZ 85212

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Office of Naval Research
Human Systems Department
Code 342 Rm 1051
875 NORTH RANDOLPH ST
ARLINGTON VA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report documents a 30-month effort sponsored by the Office of Naval Research that refined, applied and evaluated methods for analyzing the communication flow and content surrounding collaboration. Communication analysis methods were applied to the communication data from two studies in the context of a three-person Unmanned Aerial Vehicle ground control simulation and were evaluated in terms of their ability to predict team performance in a consistent manner across studies. All methods, with the exception of the Process Surrogate flow-based method, were validated by these criteria. Barriers to full automation of the methods and generalization to different domains were identified with proposed solutions. Application of the communication analysis methods revealed that high performing teams developed stable, consistent patterns of communicating which could be contrasted to teams that were distributed, under high workload, or facing a communication malfunction which were characterized by variable, yet flexible and adaptive communication patterns.

15. SUBJECT TERMS

Collaboration, teams, communication, latent semantic analysis, communication flow, team situation awareness, unmanned aerial vehicles

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE	UL		Nancy J. Cooke
					19b. TELEPHONE NUMBER *(Include area code)* 480-988-2173

USING TEAM COMMUNICATION TO UNDERSTAND TEAM COGNITION IN DISTRIBUTED VS. CO-LOCATED MISSION ENVIRONMENTS

Nancy J. Cooke, Jamie C. Gorman, Preston A. Kiekel, Peter Foltz, and Melanie Martin

FINAL PERFORMANCE REPORT

31 October 2005

ONR Grant N00014-03-1-0580

Performance Period: 24 March 2003 – 30 September 2005

Contact Information:

Nancy J. Cooke, Ph.D.
Applied Psychology
Arizona State University
7001 E. Williams Field Rd. Bldg. 140
Mesa, AZ 85212

Email: ncooke@asu.edu
Web Sites: www.certt.com
http://www.east.asu.edu/ecollege/appliedpsych/people/ncooke.html
Phone: 480-988-2173
Fax: 480-988-3162

TABLE OF CONTENTS

List of Figures

List of Tables

1.0 EXECUTIVE SUMMARY

This report documents a 30-month effort sponsored by the Office of Naval Research that refined, applied and evaluated methods for analyzing the communication flow and content surrounding collaboration. The methods include four measures of communication content based on Latent Semantic Analysis and five methods that extract patterns in communication flow.

Communication analysis methods were applied to the communication data from two studies in the context of a three-person Unmanned Aerial Vehicle ground control simulation. In the studies workload and geographic dispersion were manipulated and team performance, process, team situation awareness, and shared mental models were measured.

Communication analysis methods were evaluated in terms of their ability to predict team performance in a consistent manner across studies. All methods, with the exception of the Process Surrogate flow-based method, were validated by these criteria. Barriers to full automation of the methods and generalization to different domains were identified with proposed solutions.

Application of the communication analysis methods revealed that high performing teams developed stable, consistent patterns of communicating which could be contrasted to teams that were distributed, under high workload, or facing a communication malfunction which were characterized by variable, yet flexible and adaptive communication patterns. Findings led to an ecological perspective on team cognition, as well as new methods for assessing team situation awareness and team coordination that are inspired by this perspective. The methods can be applied to better understand collaboration or to assess collaboration in order to evaluate tools or techniques purported to enhance collaboration. With full automation and application to a wider array of domains these methods can be applied to real-time monitoring of team communication for just-in-time intervention.

2.0 RESEARCH TEAM

PI: Nancy J. Cooke
Applied Psychology
Arizona State University

CoPI: Peter W. Foltz
Psychology
New Mexico State University

Graduate Students
New Mexico State University
Preston A. Kiekel (psychology), Jamie C. Gorman (psychology), Melanie J. Martin (computer science),
Ahmed Abdelali (computer science)

Transcribers
New Mexico State University
Susan Smith, Jessica Cox, Katrina Garret
New Mexico State University

Transcribers
Arizona State University
Christy Caballero, Kendall Gans (ASU transcribers)

Undergraduate Research Assistants
Arizona State University
Sandee Scott, Amanda Taylor

CERTT Lab developer
US Positioning
Steven M. Shope

3.0 INTRODUCTION

3.1 The Problem

Technological developments in the workplace and elsewhere have drastically changed the nature of many tasks (Howell & Cooke, 1989), so that they have a much stronger cognitive component. Taking a cognitive engineering perspective, these cognitive factors need to be examined in the context of the larger sociotechnical system in which they are embedded (Hutchins, 1995; Norman, 1986, Woods & Roth, 1988). The growing complexity of tasks frequently surpasses the cognitive capabilities of individuals and thus, necessitates a collaborative approach. This is true in both military (Salas, Cannon-Bowers, Church-Payne, & Smith-Jentsch, 1998) and civilian environments (e.g., Sundstrom, DeMeuse, & Futrell, 1990).

Whereas the collaborative approach is often seen as a solution to cognitively complex tasks, it also introduces an additional layer of cognitive requirements that are associated with the demands of collaboration. An understanding of collaborative cognition, or the new "social cognition" (Klimoski & Mohammed, 1994), is critical to understanding collaborative performance. This is especially true of certain dynamic aspects of collaborative cognition, such as coordination and communication. Tasks requiring collaborative cognition in military settings are frequently embedded in a complex data-rich environment. Individuals who work together are often separated by space and time with asynchronous and distributed command-level decision making becoming the norm, rather than the exception. Participants in this decision making setting are typically of multidisciplinary, multicultural, or heterogeneous backgrounds. Further, group membership can be dynamic with the identity of participants and tasks changing over time and with high stress, high stakes, and fast tempo being common constraints. All of these factors further complicate the setting for collaborative cognition.

For collaborative cognition, as in other contexts, measurement is a critical initial step. Adequate measures are required for assessment and diagnosis of collaborative performance, and for evaluating the success of interventions to aid collaboration. The long-term goal of our research program is to develop and evaluate measures of collaborative cognition and performance. Current measures of collaborative cognition are limited given the complex nature of the typical tasks described previously (see Cooke, Salas, Cannon-Bowers, & Stout, 2000). These limitations need to be addressed in order to soundly measure and ultimately understand, team cognition.

One such limitation has to do with the fact that measures are often cumbersome and administered apart from the task. By using communication data generated relatively effortlessly as a byproduct of group interaction, we hope to overcome some of these limitations. Just as individual cognition is reflected in the behavior of the individual, collaborative cognition is reflected in the behavior of the group. Communication is one salient aspect of group behavior that is particularly tied to collaborative cognition and that has been used to infer collaborative cognition in several of the aforementioned

studies. We take this a step further and propose that *communication is cognitive processing* at the team or group level.

Information regarding the sequential patterns of communication and the flow of communication among team members (Bowers, Jentsch, Salas, & Braun, 1998) is critical to the assessment of collaborative cognition. In general, efforts in this area are hampered by the paucity of methods and tools for measuring communication in a cost-effective way (i.e., automated analyses, task-embedded, while exploiting its richness).

Before addressing our approach to this problem, we provide some background on theories, methods, and empirical findings relevant to communication.

3.2 Theories and Models of Communication

Theories and models of communication have evolved in their consideration of the continuous, complex, and directed nature of communication. For instance, early communication models were based on physical processes, such as electric current (Beebe & Masterson, 1997). Communication was thought to serve the function of reducing uncertainty, by sending information to the appropriate receivers in a mathematically defined stimulus-response pattern. For example, in the Shannon-Weaver model, information goes from the sender through a channel containing noise, to a receiver. Theories such as this one have been applied to communication, though with recognition of the interchangeable roles of sender and receiver in typical discourse (Smith, 1994, Wegner, 1995).

Further, within communication networks that arise from groups of communicators, most communication is not to the group, but to specific members of the group (Beebe & Masterson, 1997). For example, individuals may direct most of their comments to their nearest neighbors, or to the group's facilitator. Once a group has developed a communication pattern, they tend to stick to it. Groups with more equal communication structures tend to take longer to generate decisions, but the decisions tend to be more accurate. The transmission of private information to the group through artifacts and through transient communication generates a sort of group cognition and group memory (Smith, 1994). However, it is inappropriate to try to communicate all personally held information to the group, because one virtue of groups is that everyone need not know everything about every task (Smith, 1994, Wegner, 1995).

Communication theories also consider the richness of the context surrounding communication. Factors thought to influence communication include culture (Merrit & Helmreich, 1996), context of the communication, the size of the group or team, and group identity (Beebe & Masterson, 1997). As an example of a theory that incorporates these factors, structuration theory describes these and other factors as goals and conditions, which are then used to create rules that the group evaluates with respect to goals, and revises as necessary. Symbolic convergence theory focuses on the development of a common identity for the group, by engaging in mutually fulfilling social interactions.

The increasing complexity of communication theories and models has led to a need for communication analysis methods that capitalize on and make sense of the richness of communication data.

3.3 Traditional Methods for the Analysis of Communication

Most commonly, analyses of communication data have either focused on low-level quantitative measures, such as duration of communication, or on encoding the communication into prescribed content categories (Contractor & Grant, 1996). The former approach can be used to capture some of the complexity of communication patterns through time by modeling the quantitative measures using lag sequential and/or Markov chains, time series modeling, Fourier analysis (Watt & VanLear, 1996, p. 12) or other methods (Sanderson & Fisher, 1994). We refer to such data and analysis as *communication flow*.

The other common approach to communication analysis involves first selecting a coding scheme that includes all interesting categories of communication meaning, such as the rules being displayed in the conversation, the types of speech, or the actual meaning of the discussion. The transcribed discourse is then divided into the smallest units of meaning, then those pieces of text that correspond to the categories of interest are tagged (Emmert & Barker 1989). Communication patterns can be analyzed either as frequency counts of the categories or as a series of events (called "interaction analysis", for discussion, see Emmert, 1989; for an example, see Poole, Holmes, Watson, & DeSanctis, 1993), using lag sequential analysis or other tools (see Holmes, 1997 for an example). We refer to these data and analyses as *communication content*.

Both flow and content approaches have their own merit, and their own costs. For the content approach, multiple coders are intensively trained, and must have adequate agreement. Emmert and Barker (1989) cite an example of a study requiring 28 hours of transcription and encoding for each hour of communication (p. 244). But the advantage is that interpretable qualitative data are captured, including, in some cases, nonverbal communication (Donaghy, 1989). Flow approaches are much easier in data collection (although speaker, listener, and communication duration is often tedious to transcribe from audio tape), but fail to explicitly capture semantics. Both approaches have been used to analyze communication among groups larger than two, but the transcription and encoding tasks become even more cumbersome as the complexity of the communication and the possibility for parallel discourse streams increases. Content techniques are especially prone to these difficulties.

In summary, there is a general consensus that continuous streams of rich data are necessary to describe the unfolding process of communication, but that automatic methods for doing this are sparse or problematic (Smith, 1994). If researchers are interested in modeling the flow of who talks to whom and for how long, human raters must record and time-stamp these data. Content is even more labor intensive, since it requires that human raters first transcribe, then classify the discourse into prescribed categories.

3.4 Team Communication: Empirical Findings

Parallel to the methods used to analyze more general communication, team or collaborative communication can be defined by flow (e.g. Oser, Prince, Morgan, & Simpson, 1991) and content (e.g., content codes). In terms of flow, results from static measures have been equivocal. In some cases studied, high performing teams communicate with higher overall frequency than low performing teams (Foushee & Manos, 1981; Mosier & Chidester, 1991; Orasanu, 1990), but in other cases this finding has not been supported (e.g., Thornton, 1992). Some studies indicate that overall communication frequency is reduced during high workload periods (Kleinman & Serfaty, 1989; Oser, et al., 1991), whereas others indicate increases in communication frequency under relatively high workload (e.g., Stout, 1995). Some of these differences may be due to other factors such as the task or the nature of the teams. For example, Bowers, Urban, and Morgan (1992) found that the correlation between communication frequency and team performance was tied to whether the team was hierarchical or not in structure. In other cases, mixed results may be due to the use of static flow measures, devoid of semantic content or sequential information.

Communication content associated with team studies has been analyzed by segmenting transcripts into units associated with speech turns or complete thoughts. Then the segmented transcript is coded using categories pertinent to the hypothesis or research problem. Some examples of content categories include speech acts such as acknowledgments, requests, statements, or answers to questions; errors such as violation in standard format; and use of terminology such as standard military terms. Results tend to be more specific (but perhaps less generalizable) than those associated with flow analyses. For instance, Achille, Schulze, and Schmidt-Nielsen (1995) found that the use of military terms, acknowledgments, and identification statements increased with experience. Similarly, Jentsch, Sellin-Wolters, Bowers, and Salas (1995) found that faster teams made more leadership statements and more observations about the environment than slower teams.

Also parallel to general trends in communication analysis, advances in team communication analysis and understanding may come from extending analysis beyond single dimensions such as frequency of content category to more complex patterns, taking into account multiple dimensions including content, frequency, sequence, and communication flow. For instance, Bowers et al. (1998) analyzed the sequence of content categories occurring in communication in a flight simulator task. They found that high team effectiveness was associated with consistent responding to uncertainty, planning, and fact statements with acknowledgments and responses, in comparison to lower performing teams. Similarly, Bowers, Braun, and Kline (1994) found that a two-category sequence was superior to simple frequencies at predicting performance on an aerial reconnaissance task. On the basis of results like these, Salas, Bowers, and Cannon-Bowers (1995) conclude "It is likely that additional pattern-based analyses will emerge in future literature as a means to understand the impact of communication on team performance" (p. 64).

In summary, research on team or collaborative communication can focus on flow or content, and on sequential or static data. The most promising methods are sequential (either flow or content). Though much more difficult to collect, content methods do provide more specific, qualitative results than flow methods. A major obstacle in this kind of research is the costliness of manual analysis needed to transcribe and code content, and to analyze sequential flow or content data. Salas et al. (1995), highlight this research need and state, "... methods to interpret team process information, which until now has been almost exclusively a manual task, would benefit from automation" (p. 69). Indeed, collaborative cognition work in general is hampered by the paucity of automated methods and data collection limits. The methods that we will apply in this project take advantage of the richness of collaborative communication data, but are at least partially automated, making these methods more practical for the assessment of collaborative cognition.

4.0 OBJECTIVES

The overall objective of this project is to apply communication analysis methods to data sets collected in two experiments in a three-person ground control simulation of a UAV. Further, the setting was either distributed or co-located and workload varied. The communications methods applied here are semi-automated and more cost-effective than traditional manual methods and should ultimately facilitate the meaningful analysis of an extremely rich source of data on teams. Data resulting from these analyses provide a) information on the impact of factors such as distributed environments, high workload, and cognitive differences among teams (e.g., team situation awareness, shared mental models) on team communication and performance and b) methods for further exploiting communication data as an index of team performance.

This reported effort capitalizes on five specific capabilities of our research team and facility:

1) The experimental environment of the CERTT Lab offers a realistic command and control team task and provides a rich set of measures relevant to team performance and cognition.
2) Data collected in Experiments 3 and 4 in the CERTT-UAV environment specifically provide results from measures of team performance and cognition, as well as communication data.
3) Previous research has been dedicated to identifying valid measures of team performance, team process, team knowledge, and team situation awareness in this context. The reported work will leverage off of these developed metrics to test the new communication metrics.
4) The CERTT Lab's communications hardware and software automatically captures (at designated intervals) the communication flow that occurs in either direction between all pairs of individuals on a team. In addition, our research team has the expertise to apply Latent Semantic Analysis to the text generated in the course of communication.

5) The set of communication analysis methods developed and evaluated under the efforts of the previous ONR grant (see next section).

Our main objective can be divided into seven discrete tasks, not including reporting tasks:

1. ***Apply Communication Analysis Methods:*** Complete transcription of communication data from Experiments 3 and 4 and apply methods identified in previous work as having the most promise to these data.
2. ***Examine Method Validity:*** Do the analytic methods generate communication patterns that are predictive of team performance? Do the results obtained in the new experiments correspond to those of our initial study? Do they correspond to communication findings in other similar studies? Are methodological refinements indicated by the data?
3. ***Examine the Team Performance-Communication Relationship:*** How do communication patterns map onto team performance? Can we make any general statements about the communication patterns of effective or ineffective teams?
4. ***Investigate Co-located (F2F) vs. Distributed Collaboration:*** How do communication patterns change in distributed (vs. F2F/face-to-face) environments? Do these changes correspond to performance changes? Is there evidence of team adaptation through communication?
5. ***Examine Impact of Workload on Communication and Performance:*** How do communication patterns change with increasing workload and with associated performance decrements? Do teams adapt communication to changing workload? How does communication change when faced with a communication breakdown (see glitch description below)? Is the impact of the communication breakdown moderated by the environment (i.e. distributed vs. F2F)?
6. ***Investigate Link Between Communication and Shared Mental Models:*** How does the nature of the environment (i.e. distributed vs. F2F) affect the development of shared mental models or shared knowledge structures and how does communication relate to shared mental model development?
7. ***Examine Relation Between Communication and Team Situation Awareness:*** How does the nature of the environment (i.e. distributed vs. F2F) affect the development of team situation awareness and how does communication relate to the development of team situation awareness?

5.0 METHODOLOGICAL BACKGROUND

This effort relies on the data collected from teams in the context of the CERTT-UAV task using multiple measures of team performance, process, and cognition. In this section, we first describe the CERTT-UAV task. We follow this with a description of the measures used in this context, starting with the communication measures and analytic techniques that are central to this effort.

5.1 The CERTT-UAV Task

In 1998, we designed and developed a synthetic team task environment (CERTT UAV-STE) that is an abstraction of the Predator Uninhabited Air Vehicle operations (Cooke, Rivera, Shope & Caukwell, 1999; Cooke & Shope, 2002; Cooke, Shope, & Rivera, 2000). CERTT's UAV-STE is a three-team member task in which each team member is provided with distinct, though overlapping, training; has unique, yet interdependent roles; and is presented with different and overlapping information during the mission. The overall goal is to fly the UAV to designated target areas and to take acceptable photos at these areas.

The AVO (Air Vehicle Operator) controls airspeed, heading, and altitude, and monitors UAV systems. The PLO (Payload Operator) adjusts camera settings, takes photos, and monitors the camera equipment. The DEMPC (Data Exploitation, Mission Planning and Communication Operator) oversees the mission and determines flight paths under various constraints. To complete the mission, the team members must share information with one another and work in a coordinated fashion. Most communication is via microphones and headsets, although some involves computer messaging.

The CERTT UAV-STE was abstracted from results of a cognitive task analysis (Gugerty, DeBoom, Walker, & Burns, 1999) of the Predator operational environment, with the goal of providing an experimenter-friendly test-bed for the study of team cognition. As a result, cognitive aspects of the task are emphasized and other task components (e.g., the specific interface, stick-and-rudder control) have been omitted. For instance, alterations in the interface enable individual team members to rapidly acquire (within 1.5 hours) the skills and knowledge needed to work as an integral part of the team.

5.2 Communication Measurement and Analysis

The communication analysis methods that we have developed and applied take two main forms: 1) methods that focus on the content of communications, and 2) methods that focus on communication flow between team members. The content of communication (i.e., the discourse among team members and the experimenters) is recorded on digital audio tape and video tape and later transcribed by humans to generate a text file. Latent Semantic Analysis (LSA) is then applied to this file. The communication flow data are collected relatively automatically using a time-stamped intercom system resident in the CERTT Lab, which is described more fully below. Flow methods are then applied to the data logged from this system.

5.2.1 Latent Semantic Analysis

Latent Semantic Analysis is a fully automatic mathematical/statistical technique for extracting and inferring relations of expected contextual usage of words in passages of discourse. It is not a traditional natural language processing or artificial intelligence program; it uses no humanly constructed dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like, and takes as its input

only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs.

The primary assumption of LSA is that there is some underlying or "latent" structure in the pattern of word usage across contexts (e.g. paragraphs or sentences within texts), and that statistical techniques can be used to estimate this latent structure. Through an analysis of the associations among words and contexts, the method produces a high-dimensional representation in which words that are used in similar contexts will be represented as being more semantically associated. Using this representation, words, sentences, or larger units of text may be compared against each other in order to determine their semantic relatedness or additionally assessed for magnitude or salience within the high-dimensional space. A brief overview of the technical approach to applying LSA will be described here. Additional details may be found in Berry (1992), Deerwester, Dumais, Furnas, Landauer, and Harshman (1990), Landauer and Dumais (1997), and Landauer, Foltz, and Laham (1998).

In order to analyze a text or texts, LSA first generates a matrix of occurrences of each word in each context (e.g., sentences or paragraphs). In this pre-processing stage, each cell of the matrix contains a transformation of the frequency of the occurrences of each word. The transformation typically used is the log of the frequency of the word times the entropy of its frequency across all contexts. Transforms of this or similar kinds have long been known to provide marked improvement in information retrieval (Harman, 1986), and have been found important in several applications of LSA. The transforms are important for correctly representing a passage as a combination of the words it contains because they emphasize specific meaning-bearing words.

LSA then applies singular-value decomposition (SVD), a form of factor analysis, or more properly the mathematical generalization of which factor analysis is a special case. The SVD scaling decomposes the word-by-context matrix into a set of k, typically 100 to 300, orthogonal factors (or dimensions) from which the original matrix can be approximated by linear combination. Instead of representing contexts and terms directly as vectors of independent words, LSA represents them as continuous values on each of the k orthogonal indexing dimensions derived from the SVD analysis. Since the number of factors or dimensions is much smaller than the number of unique terms, words will not be independent. For example, if two terms are used in similar contexts, they will have similar vectors in the reduced-dimensional LSA representation. One advantage of this approach is that matching can be done between two pieces of textual information, even if they have no words in common.

One can interpret the analysis performed by SVD geometrically. The result of the SVD is a k-dimensional vector space containing a vector for each term and each document. The location of term vectors reflects the correlations in their usage across documents. Similarly, the location of document vectors reflects correlations in the terms used in the documents. In this space, the cosine or dot product between vectors corresponds to their estimated semantic similarity. Thus, by determining the vectors of two pieces of textual information, we can determine the semantic similarity between them. Additionally, the

geometric interpretation provides for the assessment of the domain salience of textual information via the vector length, or norm, within the vector space.

The number of dimensions retained in LSA is an empirical issue. Because the underlying principle is that the original data *should not* be perfectly regenerated but, rather, an optimal dimensionality should be found that will cause correct induction of underlying relations, the customary factor-analytic approach of choosing a dimensionality that most parsimoniously represents the true variance of the original data is not appropriate. Instead some external criterion of validity is sought, such as the performance on a synonym test or prediction of the missing words in passages if some portion is deleted in forming the initial matrix.

LSA's performance has been evaluated as a representational model and measure of human verbal concepts and has been used for a wide variety of applications that require the analysis of the conceptual content of textual material. LSA's performance has been assessed in several ways: (1) as a predictor of query-document topic similarity judgments in information retrieval (Deerwester et al., 1990); (2) as a simulation of agreed upon word-word relations and of human vocabulary test synonym judgments (Landauer & Dumais, 1997), (3) as a simulation of human choices on subject-matter multiple choice tests, (4) as a predictor of text coherence and resulting comprehension (Foltz, Kintsch & Landauer, 1998), (5) as a simulation of word-word and passage-word relations found in lexical priming experiments (Landauer & Dumais, 1997), (6) as a predictor of subjective ratings of text properties, i.e. grades assigned to essays (Foltz, 1996; Foltz, Laham & Landauer, 1999; Rehder, Schreiner, Wolfe, Laham, Landauer, & Kintsch,1998), (7) as a predictor of appropriate matches of instructional text to learners essays (Wolfe, Schreiner, Rehder, Laham, Foltz, Kintsch, & Landauer, 1998), and (8) as a predictor of team communication performance (Kiekel, Cooke, Foltz, and Shope, 2001).

While assessing the performance of LSA, the above tests also permit the derivation of applications that incorporate LSA for measuring the conceptual content of textual information. Existing applications have included information retrieval and filtering programs, techniques for automatically scoring and commenting essays, methods determining the appropriate training material for individual learners, and methods for analyzing discourse between patients and therapists. In this project, similar approaches are employed in order to analyze and categorize the discourse of team communication.

Specifically, there are four LSA-based metrics that have been developed and favorably evaluated (under prior ONR funding) in the context of a separate data set (also taken from the UAV task context). These metrics will be applied to the two data sets from Experiments 3 and 4, respectively. The metrics include: 1) LSA-based performance scoring, 2) lag coherence, 3) communication density, and 4) automatic tagging. We can also predict performance from *LSA-based measures that score performance* based on previous scores. Using a method similar to that used to score essays with LSA (Landauer, et al. 1998), we used the transcripts to predict the team performance score. We generate the predicted team performance scores was as follows: Given a subset of transcripts, S, with known performance scores, and a transcript, t, with unknown performance score, we

can estimate the performance score for t by computing its similarity to each transcript in S. The similarity between any two transcripts is measured by the cosine between the transcript vectors in the UAV-Corpus semantic space. To compute the estimated score for t, we take the average of the performance scores of the 10 closest transcripts in S, weighted by cosines (this works also if the other missions for the given team are excluded from consideration). *Lag coherence* measures are predictive of performance. The procedure is as follows: a) Take the cosine between each utterance vector and its sequel, b) Average the cosines over various lags (e.g., 36 lag moving window). c) Calculate regression equation predicting log of the cosines from log of the lag for each team-at-mission, and d) Take slope estimate of this as a measure of topic shifting or mission coherence. *Communication density* is based on the total sum vector length of all utterances in a given team-at-mission transcript and the number of words contained in the same. It thereby, extends word count to account for the amount of meaningful content being expressed by team members. *Automatic tagging* ultimately allows content category (e.g., acknowledgement, question) to be coded automatically. The procedure is as follows: a) For each utterance within a team-at-mission transcript, find the most semantically similar utterances in other transcripts that have already been tagged, b) Assign a probability of tags to that utterance. Overall, these four LSA-based metrics provide a means for automatically understanding the content, amount, and quality of information being conveyed by team members, individually or as a whole.

5.2.2 Communication Flow Data

The CERTT (Cognitive Engineering Research on Team Tasks) facility has unique and specific capabilities to support the automated recording of communication flow data. That is, team participants (up to four plus an experimenter) communicate with one another over military aviation headsets with microphones. The noise isolating properties of the headphones along with the use of noise-canceling microphones makes it nearly impossible to hear extraneous noise. The audio isolation and the physical shielding provided in the design of the consoles results in the participants becoming rapidly immersed in a task. Furthermore, the audio and physical isolation provides a strong incentive for all participants to communicate through the headsets.

Figure 1. Headset

The digital communications system is quite advanced and highly flexible. The system design allows a talker (who initiates a communications episode) to select a listener or a group of listeners. The talker initiates communications by pushing and holding down a push-to-talk (PTT) button. All communications are designed as simplex. In other words, when A is talking to B, B cannot automatically talk to A without B pushing A's PTT button. We treat A talking to B as a distinct event as compared to B talking to A. The system is additive for incoming communications traffic. For example if A is talking to B and C also begins talking to B, then B hears a mixed audio signal composed of A and C

added together. The system allows for simultaneous networked communications. For example, A can talk to B and C while at the same time D is talking to A and B.

Additional features of the system allow the participants to be listening over the headsets to the computer audio output, which includes alarms, warnings, and other audio clues built into the task scenario. Communications traffic, when present, over-rides the computer audio. The computer audio will return when the communications to a participant ceases. The system allows audio noise including static, random noise, pre-recorded distracting noise such as jet engine sounds, and non-relevant communications to be added to a particular communications link. This has the effect of making some links less attractive to use than others. Furthermore, we can completely disable any individual link in the system with the throw of a switch. For example we can allow C to talk to D but disallow D's talking to C (in Experiment 3 we disabled communications between the DEMPC and AVO for five minutes).

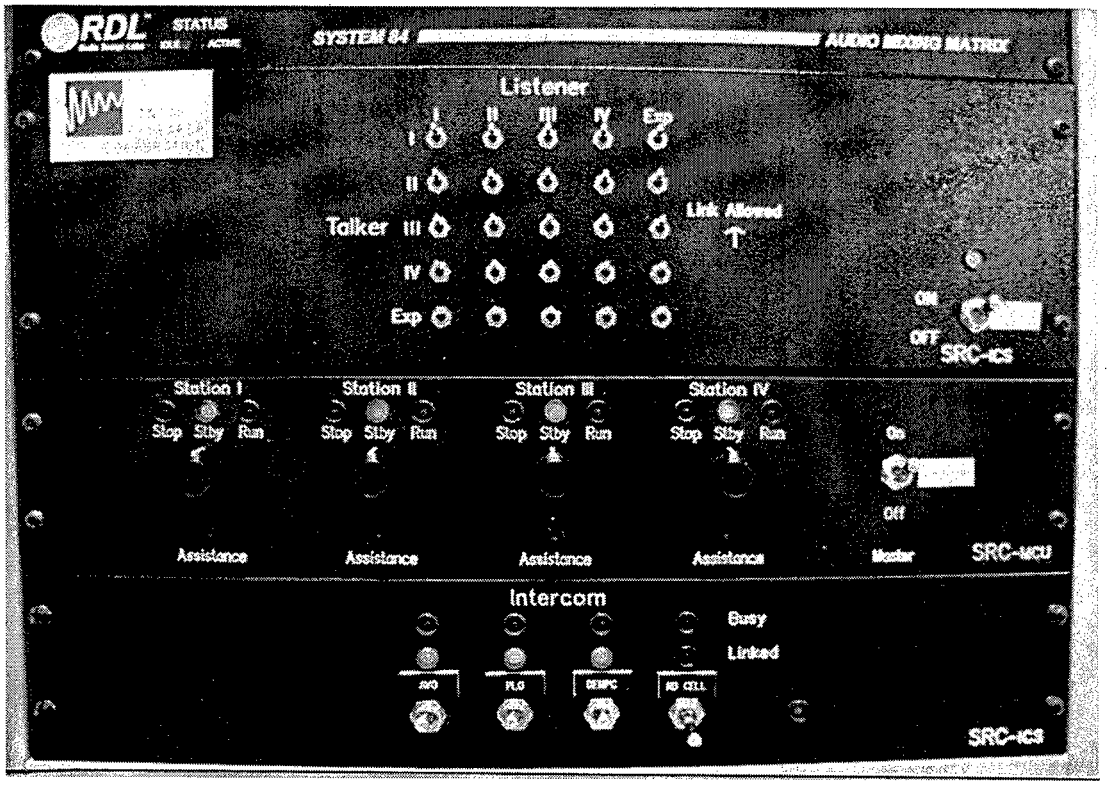


Figure 2. Experimenter's Communications Control Panel

The headset microphone output for each participant is recorded continuously, even when the PTT button is not depressed. This allows spurious individual utterances and talk-aloud statements to be recorded in addition to the intentional communication episodes. We record this microphone data on an 8-channel digital audio tape deck using a 48 kHz sampling rate. These recordings are of digital production quality. We also generate a mixed composite of all 8 channels and record this on the audio track of our video camera recorder. Two or more conversations or a message with an ambiguous sender can be disambiguated using this 8-track feature.

Most relevant to the communication flow problem, however, is the CERTT Lab's capability to record the precise timing and duration of messages from specific senders to specific receivers. In addition to the digital nature of the entire communications system, this capability is the result of two other features:

1) A push-to-talk button that must be depressed for the duration in which a message is being sent in order to be heard by the listener, and
2) Separate push-to-talk buttons are associated with each listener. So, for example, to talk to one team member, one PTT button is depressed. To talk to two team members and the experimenter, three PTT buttons are simultaneously depressed.

Our custom communication logging software samples the positions of the push-to-talk buttons at a user-selectable sample rate. We initially used a 1 Hz sample rate. However, during analysis of the COMLOG/video tape recordings in the previous ONR effort, we audibly detected some voice communication episodes on tape that were not depicted in the COMLOG data. We found that some significant communication episodes can be less than one second in length. To eliminate this problem, we have now moved to a higher sampling rate of 8 Hz (1/8 second sampling interval) reflected in the data of Experiment 4.

An $n \times n$ matrix is used to represents the state of the communication network made up of n team members. The rows represent senders and the columns, receivers. At each sample interval, we record a snapshot of this matrix. This matrix represents all possible states of the communications network, including asymmetric or directed communications. The link-disable feature described above is reflected in this matrix by certain elements always being in the "off" state.

This automated collection of communication flow data (who is talking to whom and for how long), enables investigators to bypass the manual step of transcription and coding the transcript in terms of the sender, receiver, and duration of the message. The resolution is also much greater than possible with human observers. For example, at an 8 Hz sample rate, we record the 25-element matrix eight times a second. In a one-hour task, this represents 720,000 pieces of distinct communications data.

This automation is not only cost-effective in terms of eliminating tedious transcription and encoding steps, but it also assures virtually error-free data without the need for human judgments about the identity of senders, receivers, or timing of the messages. Further, human judgments regarding the identity of the listener are virtually impossible to discern. Thus, the data collected by the communication logger can allow for rapid sequential and pattern analyses on large communication data sets. Questions concerning overall communication frequency, sender-receiver communication patterns, and various sequential dependencies can be readily answered with these data. This leaves more time to be spent analyzing (e.g., lag sequential analysis, log linear analysis, ProNet (Cooke,

Neville, & Rowe, 1996; Cooke & Gillan, 1999; Gillan & Cooke, in press)) the communication flow data.

The analysis of communication flow is critical to understanding team communication. To understand team cognition such as team situation awareness and shared mental models, the issue of not only what information is passed among team members, but also how that information is passed (i.e., to whom and by whom and at what time) is critical. To maintain team situation awareness for instance, the right information has to make it to the right individual at the right time. Information content, therefore, is only a part of the picture, and a very small one without timely delivery of that content to the team member who needs it. The CERTT data collection capabilities, combined with adequate pattern and sequential analyses, should help shed light on these types of issues.

Five methods to analyze the communication flow data have been developed and favorably evaluated (under a previous ONR-funded effort) in the CERTT-UAV context using a separate data set. They include: 1) dominance, 2) flow quantity (CRP), 3) flow sequence (ProNet), 4) stability (CHUMS), 5) flow as a team process surrogate. This suite of five methods is called FAUCET (Flow Analysis of Utterance Communication Events for Teams).

We can predict team performance from *dominance* based on very simple communication log measures. The calculation of the dominance statistic is easily automated and involves looking at the cross correlations among team members of speech quantity. The result is a ratio that indicates the degree of influence that one team member has over the others in regard to speech quantity. Specifically dominance is calculated as in Table 1.

Table 1
Dominance Metric

- Start with a time series of mean speech quantity for each team member over some small number of seconds. The number of seconds in an average is taken to be either 1/2 the mean speech duration, or 5 sec (whichever is longer).
- Take all pairwise cross-correlation functions between all team members.
- For each team member predicting each other team member, take the weighted average of the cross-correlation function, where the weight is the inverse of the lag.
- For each team member predicting each other team member, take the ratio of one mean cross-correlation to the other (i.e. person A to B's correlation is divided by the correlation for B to A).
- Take the natural log of all of these ratios. Because of the properties of the correlation coefficient, this ratio will be approximately normally distributed, with a mean of zero and a standard deviation of 1.
- For each team member, the pair of log ratios can be averaged, yielding a mean score of influence that the team member exhibits over other team members.

Flow quantity is simply a metric representing the amount of speech to and from each team member. CRP (Communication Required and Passed) scores measure how far a team deviates from an ideal in terms of relative speech ratios, across the whole mission. Ideal speech ratios indicate how much each team member should speak to each other team member. CRP analyses yield a score for each team member, which is itself an aggregate of two components ("chat" and "information" components). Low scores (approaching 0) indicate high deviance from this overall ratio, and high scores (approaching 1) indicate that the team is speaking in approximately ideal ratios. Specifically, CRP is calculated as in Table 2.

Table 2
CRP Metric

- Start with a separate sum of every second when team member X is talking to team member Y. Let us define the total sum of seconds that x spends speaking to some other person y as C_{xy}. Create a proportion, relative to all possible seconds.
- For each team member, take the proportion of time that that person is either speaking or being spoken to. Take the converse of this proportion, to yield a positive correlation with performance. This is the "Chatter" score.
- Define the minimal amount of speech each team member must convey to each other team member to complete the mission. For team member X talking to Y, this can be defined by the number of sentence clauses X must convey (or request) from Y (b_{xy}), times the number of events requiring this transmission (U_{xy}), times an arbitrary constant of how long it takes to convey a single clause (k).
- Ideally, $C_{xy} = U_{xy}(b_{xy}k)$for all persons y, x.
- Therefore, ideally, for person x, $C_{yx}/C_{zx} = [U_{yx}*(b_{yx}*k)]/[U_{zx}*(b_{zx}*k)]$
- Therefore, ideally, $C_{yx}/C_{zx} = [U_{yx}*b_{yx}]/[U_{zx}*b_{zx}]$
- Therefore, ideally, $\{C_{yx}/C_{zx}\} / \{[U_{yx}*b_{yx}]/[U_{zx}*b_{zx}]\} = \{C_{yx}*U_{zx}*b_{zx}\}/\{C_{zx}*U_{yx}*b_{yx}\} = 1$
- Take $\min(\{C_{yx}*U_{zx}*b_{zx}\}/\{C_{zx}*U_{yx}*b_{yx}\}, 1 / [\{C_{yx}*U_{zx}*b_{zx}\}/\{C_{zx}*U_{yx}*b_{yx}\}])$ to retain a 0-1 scale.
- This is the "Information Passing" score.
- For each team member, multiply "Chatter" by "Information Passing" to get a "Communication Required and Passed" (CRP) score.

Flow sequence adapts a sequential data analysis procedure called ProNet (Cooke, Neville, & Rowe, 1996) to define representative chains of sequential patterns of speech. ProNet can be used to define representative chains of sequential patterns in the events defined by each team member beginning or ending a speech sequence. Summary statistics are taken on lengths of the set of chains remaining for each team-at-mission. Chain length is a measure of how much stability is found in the set of utterances, on average. Therefore, mean chain length, for example, is a single measure of

communication stability. Other measures include median, minimum, maximum, standard deviation, etc.

The *stability* of the communication log data is captured by a procedure called CHUMS. CHUMS models the team at discrete time intervals (in this case minutes), then aggregates the models based on model fit. In this case, the models are multinomial models of how much each team member speaks. Two variations of the model are a) categories including each person speaking, plus the "null" event (when no-one is speaking) or b) categories that do not include "null" events. CHUMS gives a single value for the team, such as models remaining after clustering, or models per minute. These statistics reveal how many statistically distinct patterns of relative speech quantity the team exhibited during the mission, and so is a measure of communication stability. Teams with more stable communication styles will exhibit fewer distinct models and models per minute. A negative correlation between process variables and number of CHUMS models indicates that teams with more stable communication also tend to score higher on process measures. Specifically CHUMS is calculated as in Table 3.

Table 3
CHUMS Metric

- First the communication log files are separated into one-minute intervals.
- For each minute, model the communication frequency of each team member, with a multinomial model.
- Use the model for each minute to test the data for each other minute, using a chi-square approximation to the multinomial test.
- Perform an iterative agglomerative cluster analysis on the minutes of the mission, using model fit as a distance measure.
- Count the number of remaining clusters.

Team performance scores can be adequately predicted from individual scores aggregated using the communication data as an *estimate of process behavior*. This work forms Preston Kiekel's M.A. project for Experimental Statistics, which was successfully defended in May of 2003. Basically, three communication-based functions of individual performance scores were defined to convert them into a commensurate form with the team performance score. The specific procedure is outlined in Table 4.

Table 4
Process surrogate

- Three communication-based functions of individual performance scores were defined to convert them into a commensurate form with the team performance score. All three consisted of first proportionalizing the individual performance score. Then said score is:
    - multiplied by the CRP score defined above,
    - multiplied by the Dominance score defined above, and
    - multiplied by a proportionalized score from an observational process check list.
- Three aggregation schemes are employed on the transformed individual scores yielded by each of these three functions. The sets of individual scores were therefore converted to single scores to be correlated with performance. A total of nine aggregates were taken, in that each of the three functions above were then aggregated by:
    - arithmetic mean
    - minimum
    - geometric mean
- The arithmetic mean of the untransformed individual scores is taken as a parsimonious baseline.

5.2.3 Team Performance Measures

Prior to the start of this effort, team performance metrics had been validated in the CERTT-UAV context. Team performance is measured in the CERTT-UAV task using a composite score based on the result of mission variables including time each individual spent in an alarm state, amount of fuel used, amount of film used, number of missed targets, number of critical waypoints missed, time spent in a warning state, and route sequence violations. Penalty points for each of these components are weighted a priori in accord with importance to the task and subtracted from a maximum score of 1000. Specifically, missed targets are weighted four times that of fuel and film used and alarm time is weighted two times fuel and film used. Critical waypoints missed and route sequence violations are weighted three times fuel and film used. Warning time is weighted the same as fuel and film used. Team performance data were collected for each 30-40-minute mission.

In the experiments of interest to this project, missions differed on the basis of number of target waypoints (that needed to be photographed). Low workload missions contained 9 targets for a 40-minute (maximum) mission and high workload missions contained 20. Thus, a team that gets photographs of nine targets in both types of missions (comparable team performance) would score lower in the high workload mission, simply because there were more targets to miss.

Therefore, modifications were made to our previous metric of team performance in order to base team performance on the rate with which tasks were completed (e.g., number of photos per minute) rather than the proportion of tasks that were completed (e.g., number of photos taken out of total possible). This revision accommodates scoring of the high workload scenario, and other variations of the mission scenarios, and prevents penalizing teams for not achieving similar proportions of outcome across different scenarios. For example, the new team performance metric, which is based on rate of performance, does not penalize teams for photographing a smaller proportion of targets in the high workload missions (e.g., 12 out of 20) despite the improvement from the low workload missions (e.g., 9 out of 9).

Furthermore, in order to make the team score more independent from the individual role scores, we removed penalties for fuel, film, and route sequence violations, as these penalties are specific to only one role. Finally, the relative weighting scheme used in the team performance and individual role performance metrics was also revised to better differentiate between team and individual tasks or components. For example, the "missed or slow photo penalty" component was given lower weight for the PLO score but higher weight for the team score, as this task requires effort on the part of all team members and is not solely the PLO's responsibility. In general, components of the individual role performance metrics were given a higher weight if those components, or tasks, were controlled solely by that role.

Each individual role within a team (AVO, PLO and DEMPC) also had a composite score based on various mission variables including time spent in alarm or warning state as well as variables that were unique to that role. Penalty points for each of the components were weighted *a priori* in accord with importance to the task and subtracted from a maximum score of 1000. The most important components for the AVO were time spent in alarm state and course deviations, for the DEMPC they were critical waypoints missed and route planning errors, and for the PLO, duplicate good photos, time spent in an alarm state, and number of bad photos were the most important components. *Individual performance* data for a role were collected for each of the seven missions.

Score results are displayed to team members after each mission. In addition to the individual and team scores for that mission, team members are able to see individual and team scores from previous missions of their team and other teams.

5.2.4 Team Process Measures

Team process metrics have been validated in the CERTT-UAV context. Team process behavior is scored independently by each of the two experimenters. In Experiments 3 and 4 for each mission the experimenters observe team behavior and respond to a series of six questions. Three of these items concern team behaviors that did or did not occur at designated event-triggers in each mission (e.g., within five minutes after the end of the mission, the team discusses and assesses their performance). These items are scored with either a 0 (not present) or 1 (present). The other three items also assess team behaviors that did or did not occur at designated event-triggers in each mission, but these items are

scored on a scale that ranges from very poor/none (0) to either good (2) or very good (3). The sum of scores on these six items is then expressed as a proportion of total possible points (10) for a given mission. This proportion forms the critical incident process score for each mission and team.

Four summary scores for each team are also used to assess team process for a given mission. Summary scores are based on experimenter judgments on four dimensions (communication and coordination, team decision-making, team situation awareness behaviors, and process behaviors) which are scored on a five-point scale that ranges from 1 (terrible) to 5 (excellent). Experimenters are aided when making their judgments by informal tallies that are kept for each dimension throughout the session.

5.2.5 Team Knowledge Measures

Team knowledge measures (including team situation awareness) have been validated in the CERTT-UAV context. For Experiments 3 and 4 we measured three forms of team knowledge: 1) team situation awareness (fleeting knowledge or situation models that the team has of the task and team at any one moment), 2) taskwork knowledge, and 3) teamwork knowledge. Situation awareness is measured during each mission. Taskwork and Teamwork knowledge are measured in two or more separate sessions set apart from the task.

Team situation awareness is measured using two SPAM-like (Durso, Hackworth, Truitt, Crutchfield, and Nikolic, 1998) queries administered at two randomly selected 5-minute intervals during each mission. One of the experimenters administers the queries to each individual in turn and then to the team as a whole. This latter query is an attempt to elicit team knowledge in a more holistic manner. Order in which individuals is queried is also random.

One of the two queries requires team members to make a prediction regarding the number of targets out of nine (or 20, depending on the level of workload) successfully photographed by the end of the mission, and the other query varies with the mission but generally requires prediction. The experimenter also records the correct responses to these queries given how the situation plays out and this key is used to score the eight responses for accuracy. Team accuracy scores are based on the sum of all individuals' accuracy scores. Responses to all queries are also scored for intrateam similarity. Team similarity is the sum of all the pairwise similarities of the three team members.

The team's knowledge of the task (i.e., team taskwork knowledge) is measured using a pairwise relatedness rating task. The taskwork ratings are made by team members on pairs of eleven task-related terms: altitude, focus, zoom, effective radius, ROZ entry, target, airspeed, shutter speed, fuel, mission time, and photos. All possible pairs of these terms are presented in one direction only, one pair at a time. Pair order is randomized and order within pairs is counterbalanced across participants. Each team member rates the relatedness of each pair on a 1-5 scale with anchors that range from slightly related to highly related. There is also an option of unrelated.

Taskwork consensus ratings consist of the same pairs as taskwork ratings (randomly presented), however the ratings are entered as a team. For each pair, the rating entered in the prior session by each team member is displayed on the computer screen of that team member. The three team members discuss each pair over their headsets until consensus is reached. Again, consensus measures are taken to elicit knowledge at the team-level.

Teamwork knowledge is measured using a questionnaire. A task scenario is first described and then each individual participant is required to indicate which of sixteen specific communications are absolutely necessary in order to achieve the scenario goal. The teamwork consensus ratings are administered in the same manner as the teamwork ratings, but are completed on a team level where team members discuss their answers over the headsets until a consensus is reached.

Taskwork and teamwork knowledge measures are scored for accuracy and intrateam similarity. Individual accuracy scores and pairwise measures of response similarity are averaged across team members. For the two rating tasks, data are first submitted to KNOT (using parameters r=inf. And q=n-1) in order to generate Pathfinder networks (Schvaneveldt, 1990). These networks reduce and represent the rating data in a meaningful way in terms of a graph structure with concept nodes standing for terms and links standing for associations between terms. A referent network generated by the experimenters serves as the key, and similarity of any one network to this referent in terms of the proportion of shared links is used as a measure of accuracy. In addition, the individual task ratings are scored not only against a key representing overall knowledge, but also against role-specific keys. In this way, measures of "role" or "positional" accuracy, as well as "interpositional" accuracy (i.e., interpositional knowledge (IPK) or knowledge of roles other than their own) can be determined. Team accuracy is the mean accuracy across team members. Intrateam similarity is measured using the proportion of shared links for all intrateam pairs of two individual networks (i.e. the mean of the three pairwise similarity values among the three networks).

5.2.6 Summary of Measures

In the proposed work we plan to leverage off of the various measures developed and validated previously in the CERTT-UAV context (see Table 5 below) as a basis for evaluating our communication metrics. All measures are taken at every mission except taskwork and teamwork knowledge which are taken in separate sessions apart from the missions. Communication is recorded continuously during each mission. Team knowledge measures are taken at the individual and team level. At the individual level knowledge measures are scored for accuracy (overall and positional) and intrateam similarity in order to provide a knowledge profile of the team. Our goal in the proposed effort is to 1) determine the validity of the communication metrics outlined in Table 5 in terms of their ability to discriminate high vs. Low performing teams and high vs. low knowledge teams. In addition, we plan to test hypotheses regarding the impact of workload and distributed mission environments on team communication and performance.

Table 5
Measures taken in Experiments 3 and 4

MEASURES	
Validated in Previous Efforts	**New Communication Metrics (require further validation)**
Team performance score (rate version)	Content – LSA-based density
Individual performance score	Content- LSA-based performance score
Team process-critical incident	Content – LSA-based automatic tagging
Team process-summary rating	Content- LSA-based lag coherence
Team knowledge-situation awareness queries	Flow - Dominance
Team knowledge – teamwork knowledge	Flow – Quantity: CRP
	Flow – Sequence: ProNet
	Flow – Stability: CHUMS
	Flow – Team process surrogate

6.0 TECHNICAL APPROACH AND RESULTS

In the following sections, we describe the data sets that we worked with and detail our technical approach associated with the seven tasks of the proposed effort.

6.1 Two CERTT-UAV Data Sets

We currently have communications data from two experiments, each in the context of a three-person Uninhabited Air Vehicle (UAV) synthetic task. This synthetic task is based on actual UAV ground operations. The goal is for the team to safely fly the UAV to targets and to orient the vehicle so that good photographs of the targets can be taken. The three team members have different roles and the mission can only be accomplished by interaction and information sharing among the three roles. Each team member is trained on unique material and has access to two unique information displays. For instance, the Air Vehicle Operator (AVO) is trained to navigate the UAV to specific waypoints and to adjust speed and altitude based on restrictions and photograph requirements. The AVO's displays and controls are centered on these navigation tasks. In addition, each operator has a unique screen to monitor in order to avoid system alarm states. This particular feature enables the manipulation of workload in the scenario. Flexible task software also allows for rapid scenario changes (e.g., target locations, hazard locations). Communication takes place over headsets with microphones. As mentioned in the previous section, communication data consists of audio captured on a digital 8-track audio recorder, audio recorded on video tape, as well as communication log data indicating sender, receiver, and message duration. In addition, other measures including task-embedded individual and team performance measures, team process measures, team

situation awareness measures, and measures of individual and team taskwork and teamwork knowledge are taken.

In Experiment 3 (data collected in fall of 2001 and spring 2002), communications data were recorded over seven consecutive 40-minute missions for 20 three-person teams of NMSU undergraduates. Half of the teams were co-located and half were distributed. Teams were randomly assigned to either a co-located or distributed condition. In the co-located condition the three team members were in view of each other during and between missions, though they communicated over headsets during missions. In the distributed condition, team members never saw each other or the screens of the other applications and communicated solely over headsets. Missions 1 through 4 were low workload scenarios with 9 targets and missions 5-7 were high workload scenarios with 20 targets and additional route constraints. During the 6th mission communications between the DEMPC and the AVO were severed for 5 minutes after the first *ad hoc* target was called in by the experimenter. Missions were performed in two sessions, both occurring within 48 hours of each other. Session 1 also included a 1.5-hour training session (factual and skill-based). Knowledge was measured in Session 1 immediately after training and again in Session 2 after the seventh mission. Performance, process, and situation awareness measures were taken, along with communication measures for each mission. In addition, other measures were taken that are not the focus of this effort. They include demographic data, leadership questionnaires, subjective measures of situation awareness and workload, and measures of individual working memory. The main goal of this experiment was to examine the effects of distributed mission environments on team skill acquisition and skilled performance.

Experiment 4 (data collected in fall 2002) was a replication of Experiment 3 with 20 all male teams. This was done in order to reduce some variation attributed to mixed gender teams in Experiment 3 mission. Experiment 4 procedures were the same as those for Experiment 3 except that all male teams were obtained, there were 5 missions (with the last one being high workload), and only one knowledge measurement session that occurred after mission 3.

The tasks below were conducted using the data sets collected in the two experiments described in the previous section.

6.2 Hypotheses

Based on the team and communication literature as well as previous results from two other CERTT-UAV experiments, we formulated hypotheses about team performance, cognition, and communication under the conditions described in the previous section. We have broken these down by task. We will discuss them in turn within each task section.

Table 6
Hypotheses associated with Tasks 2 through 7

Task/Hypothesis Number	Hypothesis
2.1	Communication metrics, if valid, should be correlated with performance scores.
2.2	Functions relating performance and communication metrics should generalize across studies.
3.1	Teams with higher performance scores will have greater disparity among communication flow dominance scores.
3.2	High-performing teams will have (according to the flow sequence analysis) more completed speech cycles and fewer interruptions than low-scoring teams.
3.3	High-performing teams should have longer mean chain lengths in the sequential flow analysis than low-performing teams.
3.4	High-performing teams will have fewer patterns of communication frequency (in terms of flow stability metric) than low performing teams.
3.5	Better teams will follow assertions or action statements with acknowledgements as based on LSA coding, more than ineffective teams.
3.6	Better teams in terms of performance will have communication efficiency scores that fall within a mid-level range.
3.7	As team skill is acquired, we predict that individual teams will start out as inefficient and then proceed to being overly "efficient" and finally reach the point of optimal efficiency.
3.8	Effective teams will demonstrate greater coherence at longer lags compared to ineffective teams who will demonstrate less coherence at longer lags.
3.9	Individually, teams should start out with small lag coherence and at asymptotic levels of performance should demonstrate greater lag coherence.
4.1	Performance-communication relationships identified under Task 3 should be stronger in the distributed condition than the co-located condition.
4.2	We expect LSA-based coding to reflect more teamwork and non-task related communications among distributed teams than co-located teams.
4.3	The communication differences found in 4.2 should dissipate with experience.
5.1	Teams will communicate less under high workload than low.
5.2	LSA-based coding will indicate more action-oriented communications under high workload compared to low.

5.3	The patterns in 5.1 and 5.2 are expected to hold more for high-performing teams than low.
5.4	High-performing teams should be faster to switch into and out of the alternative communication paths employed during communication breakdowns than low-performing teams.
5.5	Under low workload, high-performing teams should have high coherence scores, but under the later high workload missions, higher-performing teams should have lower coherence scores.
6.1	Teams with more taskwork knowledge should have longer mean chain lengths representing communication flow.
6.2	Teams with high levels of teamwork knowledge should exhibit more stability in terms of communication flow.
6.3	Increased interpositional knowledge and intrateam similarity should correspond to decreases in communication frequency.
7.1	Teams with higher levels of team situation awareness will follow statements with acknowledgements.
7.2	Differences in team situation awareness will be reflected by changes in flow patterns.

6.3 Task 1: Apply Communication Analysis Methods

6.3.1 Method

The flow methods described above (dominance, flow quantity, flow sequence, CHUMS-based stability, and flow as a process surrogate) were applied to the communication log data collected during the course of each mission. A emerged with the sampling rate for comlog data was increased between experiments. The comlog sampling rate was a technical problem for Dominance and CHUMS-based Stability measures. It was in both cases by averaging over one second, so that there was still a single value between 0 and 1 for each second. We applied all five comlog methods to Experiments 3 and 4 (ProNet, Dominance, CRP, CHUMS, and Team Process Surrogate).

In order to apply the LSA-based content analysis techniques to the Experiments 3 and 4 communications data, the audio data was first transcribed. Software that integrates our digital intercom data with the audio recordings facilitates this process and was developed as part of this effort. Videotapes are first digitized (i.e., converted to mpeg file format). Our transcription software then makes use of the communication log (speaker, listener, data) data, which is synced with the digitized video. The transcriptionist is presented with a window displaying the identity of the speaker and listener, as well as time. The transcriptionist then types what is said in the space provided. The transcription software also tags the transcripts with XML tags needed by the LSA software. These tags indicate speaker, listener, and duration information.

However, we experienced a significant bottleneck in the transcription process, resulting in a large lead for application of flow methods and a large lag for application of content methods. During this process we also identified a ComLog header – transcription interface incompatibility that resulted in errors in sequencing mission utterances. These had to be largely corrected by hand, resulting in further delays. Ultimately, these hiccups lend support to the hypothesis that automatic speech recognition would indeed be preferable to hand transcription. The finding at the end of this section that LSA measures are relatively robust to speech recognition errors in comparison to straight transcription supports this objective.

Four transcripts from Experiment 3 were not viable. Additionally, four missions from Experiment 4 were un-transcribed due to incomplete comlog files. Segments (e.g., individual utterances by team members) were compared to each other and generally assessed within a derived semantic space based on a corpus of domain-relevant information, including interviews with subject matter experts, UAV training materials, and UAV transcripts. This approach permits the assessment of the amount of semantic relatedness of utterances between team members and across communication channels. Additionally, by computing the vector lengths of utterances, we derived a measure of the density (or quality) of information being communicated in a mission. These measures therefore provided an indication of both how domain-relevant information was being communicated and how much it was being communicated. Because LSA provides a continuous measure of relatedness, the measures can be automatically converted into maps showing how the quality and relatedness of information flows among team members through their communication channels. LSA-based performance scoring, lag coherence, communication density, and automatic tagging metrics and procedures described in the methodological background section were applied to the resulting transcripts.

6.3.2 Results

Under this task the nine flow and LSA-based methods listed earlier in Table 5 were applied to the data from Experiments 3 and 4. In addition, there were several other methodological developments that were achieved under this task and applied to the same data. These include:

- Metrics to evaluate tagging agreement for LSA-based automatic tagging and automatic speech recognition case
- A keyword-based method to provide a baseline for LSA
- A Web-based LSA interface

These miscellaneous developments are described in the following results section.

6.3.2.1 LSA tagging. Our goal is to use semantic content of team dialogues to better understand and predict team performance. One approach is to look at the dialogues as a whole, which we will discuss later. The approach we focus on here is to study the dialogue on an utterance or turn level. To this end we chose the Bowers Tag Set and

manually annotated the transcripts. We then developed an algorithm to tag transcripts automatically, resulting in some decrease in performance, but a significant savings in time and resources.

The Bowers tag set. Bowers and colleagues (1998) analyzed the sequence of content categories occurring in communication in a flight simulator task. They found that high team effectiveness was associated with consistent responding to uncertainty, planning, and fact statements with acknowledgments and responses in comparison to lower performing teams. We used the same tags developed by Bowers et al. to categorize statements made by team members. A subset of the statements was manually annotated and then these annotations were compared against an automatic tagging performed by LSA. The tags developed by Bower et al. are shown in Table 7.

Table 7
Bowers Tag Set

Tag	Definition	Explanation
A	Acknowledgment	One-bit statements answering the previous statement, such as "yes," "no," "roger;" Could also follow an action.
AN	Action	Statements that require a particular crewmember to perform a specific action-- including the speaker, immediate, precise, like a command.
EXP	Experimenter	Non-task communications that were directed to or came from the experimenter.
F	Factual	Statements that verbalize readily observable realities of the environment, any objective facts (even if wrong), including statements of immediate past action ("I did this...").
NT	Non-task	Non-task related statements.
P	Planning	Statements-- not always person specific, less immediate, less specific than AN, anything relevant ONLY in the future of the mission, has to be affirmative, not a question, not just one WP into future.
Q	Unknown	Can't be tagged.
R	Response	Statement that are differed from acknowledgments only in that they conveyed more than one bit of information-- A plus more info.
U	Uncertainty	Statements which included direct and indirect questions.

Manual annotation for Experiment 1 data. Three annotators, two psychology graduate students familiar with the project and one undergraduate, each tagged 26 or 27 team-at-missions (in the Experiment 1 data set), using the Bowers' Tag Set, so that 12 team-at-missions were tagged by two annotators. Initially, we chose to measure inter-coder reliability using the C-value measure (Schvaneveldt, 1990).

The C-value was chosen for its ability to handle arbitrarily long sequences of tags for a given turn and cases where taggers assigned sequences of tags of different lengths to a given turn. Turns consist of one or more utterances by a single speaker and it is possible that each utterance could have a different tag. For example:

"DEMPC to AVO. Okay SEN1 has speed of max 200, and altitude between 3000 and 5000. You need to go a little bit to your right. Effective radius of five miles."

was tagged as "F" by one tagger and "F-AN-F" by another, where:

"F" indicates (objective) factual statements, and
"AN" indicates an action statement which
requires a crew member to perform a specific action.

The C-value for a turn is computed by taking the number of tags in the intersection of the sets of tags assigned by the taggers, divided by the union of these sets. So in our example, the two sets are {F} and {F, AN}. The intersection is {F} and the union is {F, AN}, so the C-value is 1/2 or 0.50. The C-value for a turn ranges between 0 and 1, where 0 indicates an empty intersection of the tag sets, or complete disagreement, and 1 indicates that the union and intersection contain the same number of tags, or complete agreement. Once the C-value is computed for each turn in a team-at-mission, we compute the average C-value for the team-at-mission by summing the C-values for the turns and dividing by the number of turns in the team-at-mission. Similarly, we compute the average C-value for the corpus (or any desired subset of the corpus).

The C-value for the 12 team-at-missions that were tagged by two annotators was 0.70. This value was used as the benchmark upon which to compare our automated tagging approaches.

Table 8 below shows the frequency of tags in the 12 transcripts tagged by two taggers. We can establish a baseline of tagging performance of 0.27, by noting that if all utterances were tagged with the most frequent tag, "F", our percentage correct would be 27%, as shown in Table 8.

Table 8
Tag Frequency Percentages

Tag	Percentage of Occurrences
F	27
A	24
U	17
AN	15
R	15
Other (P, Q, NT)	2

Automatic annotation with LSA. In order to test our algorithm to automatically annotate the data, we computed the "corrected tag" for all 2916 turns in the 12 team-at-missions tagged by at least two taggers. We used the union (no repetitions) of the sets of tags assigned by the taggers as the "corrected tag". For example see Table 9.

Table 9
"Corrected Tag" Example

Tagger 1	Tagger 2	Corrected Tag
F-AN	A-F-AN-F-AN	A-F-AN
F-AN	R-AN	F-R-AN
R-AN	F-R-AN	F-R-AN
R	A	R-A
R	A-AN-F	R-A-AN-F

The union, rather than the intersection was used since taggers more frequently missed relevant tags within an utterance and thus the union of multiple taggers might capture all likely tag types within the utterance.

Then, for each of the 12 team-at-mission transcripts, we automatically assigned "most probable" tags to each turn, based on the corrected tags of the "most similar" turns in the other 11 team-at-missions. For a given turn, T, the algorithm proceeds as follows:
Find the turns in the other 11 team-at-mission transcripts, whose vectors in the semantic space have the largest cosines, when compared with T's vector in the semantic space. We choose either the ones with the top n (usually top 10) cosines, or the ones whose cosines are above a certain threshold (usually 0.6). The corrected tags for these "most similar" turns are retrieved. The sum of the cosines for each tag that appears is computed and normalized to give a probability that the tag is the corrected tag. Finally, we determine the predicted tag by applying a cutoff (0.3 and 0.4 seem to produce the best results): all of the possible tags above the cutoff are chosen as the predicted tag. If no tag has a probability above the cutoff, them the single tag with the maximum probability is chosen as the predicted tag.

We also computed the average cosine similarity of T to its 10 closest tags as a measure of certainty of categorization. For example, if T is not similar to any previously categorized turns, then it would have a low certainty. This permits the flagging of turns that the algorithm is not likely to tag as reliability.

We applied the above algorithm to the transcripts with turns involving the experimenter role removed, because those interactions are generally irrelevant to team performance and might interfere with tag sequence analysis.

Finally, we computed the C-value between the tag predicted by the computer and the corrected tag. The results are shown in Table 10.

Table 10
Values for LSA Algorithm

Method	C-Value	Av. Cert.	Cut	Turns
Top 10	0.56	0.61	0.3	2916
Thresh 0.6	0.59	0.65	0.3	2507

In Table 10, "Top 10" indicates the algorithm that selects the ten most similar turns, while "Thresh 0.6" indicates the algorithm that selects similar turns where the cosine between the vectors is greater than 0.6. "Av. Cert." is the average certainty. "Cut" is the cutoff value. "Turns" is the total number of turns included in the calculations. Note that the total number of turns in the 12 team-at-mission transcripts is 2916 The number of turns considered is reduced from 2916 to 2507 when the algorithm with threshold 0.6 is used, because turns where no turn vectors have a cosine greater than 0.6 are excluded. Based on the C-Values, the two methods perform only 20% and 16% below the performance of human-human agreement. Considering that the approach only uses one measure, a semantic similarity measure, but ignores any syntactic measures, the results are quite promising.

In order to improve our results, we considered ways to incorporate simple discourse elements into our predictions. We used two discourse features:

1. For any turn with a question mark, "?", we increased to probability that uncertainty, "U", would be one of the tags in its predicted tag.
2. For any turn following a turn with a question mark, "?", we increased to probability that response, "R", would be one of the tags in its predicted tag.

We refer to our original algorithm with these two discourse features added as "LSA+" algorithm. Using LSA+ with our two methods now performs only 11% and 10% below human-human agreement, as shown in Table 11.

Table 11
Values for LSA+ Algorithm

Method	C-Value	Av. Cert.	Cut.	Turns
Top 10	0.62	0.58	0.3	2916
Thresh 0.6	0.63	0.65	0.4	2610

Thus, the results suggest that we can automatically annotate team transcripts with tags. While the approach is not quite as accurate as human taggers, LSA is able to tag an hour of transcripts in under a minute. As a comparison, it can take half an hour or longer for a trained tagger to do the same task.

Computing Cohen's Kappa. A commonly used coefficient of inter-coder agreement for discourse and dialogue studies is Cohen's Kappa (Cohen 1960), which takes into account chance agreement. Cohen's Kappa is defined:

$$K = \frac{P(o) - P(e)}{1 - P(e)}$$

Where P(o) is the proportion of agreement observed, and P(e) is the proportion of agreement expected by chance. In order to improve the comparability of our results to work in the discourse processing community we computed Kappa for our inter-coder agreement. For the observed agreement we used the C-values (computed as discussed above). Traditionally in tests of agreement, the proportion of expected agreement can be easily computed because only one tag is assigned to each utterance. In this case, since a tagger could assign one or more tags to each utterance, P(e) was estimated by running a Monte Carlo simulation.

In the Monte Carlo simulation for each of *n* iterations, we randomly choose the number of tags, *k*, for turn *i*, based on the frequency of tag length for turns in the corpus, and then randomly choose *k* tags for turn *i*, based on the frequency of tags in the corpus. We then do the same for turn *i+1* and compute the C-value of agreement between the randomly assigned tags for turn *i* and turn *i+1*. This procedure is repeated *n* times and the average C-value is computed. We ran our simulation for *n* equal to five million and the average C-value was 0.21. This approach provides an accurate estimate of expected chance agreement, P(e), under the assumption that taggers use the same frequencies of tags as those who participated in the study. The results of using P(e) = 0.21 in the formula for Kappa are shown in Table 12.

Table 12
Kappa and C Values

Coders-Agreement	C-Value	Kappa
Human-Human	0.70	0.62
LSA-Human	0.59	0.48
LSA+-Human	0.63	0.53

Issues arising with the computation and interpretation of Kappa have been discussed by many authors, including: Grove et al. (1981), Carletta (1996), Di Eugenio (2000). Of the

scales that have been proposed for interpretation of Kappa, Di Eugenio (2000) notes that the discourse processing community has generally adopted Krippendorff's (1980) fairly strict scale. It discounts any variable with $K < 0.67$ and allows tentative conclusions when $0.67 < K < 0.8$. There are other less strict scales, for example, Rietveld and van Hout (1993), which considers $0.42 < K < 0.6$ as indicating moderate agreement and $0.61 < K < 0.8$ as indicating substantial agreement. According to Grove et al. (1981), the psychiatric community considers $K > 0.6$ or $K > 0.5$ as acceptable. Based on the range of these scales and the complexity of the tagging task we believe that we have moderate agreement in the "Human-Human" and "LSA+-Human" categories

Automated speech recognition. Using data collected from another project that performed automated speech recognition (ASR) on three of the transcripts from Experiment 1, we tested LSA's ability to tag the ASR data. Rong and Rudnicky at CMU ran an experiment to determine baseline speech recognition accuracy on three of the transcripts in our corpus. We ran our LSA + Syntax algorithm to predict the tags on their output. Word error by the speech recognition system (Sphynx) was 38.9%. Loss of tagging accuracy, measured by C-value, was 14.9% when compared to using the original transcripts. Preliminary results indicate that noise introduced by current speech recognition technology may be mitigated by LSA's ability to detect semantic similarity.

6.3.2.2 Keyword analysis. We investigated alternative methods for evaluating communication content, including word counts and key word indexing (KWI), in order to begin to understand what additional power our more sophisticated (e.g., LSA, CRA) methods provided. Word counts are relatively straightforward. Essentially, the number of words-per-utterance were tallied and further aggregated to the desired degree of granularity, e.g., average words per utterance, average number of words in transcript, etc. KWI is a bit more complex, but allows us to compute vector lengths, cosines, and distances between utterances in a transcript. Similar to LSA, in KWI vector length is taken as a measure of domain relevance, cosines measure relatedness between utterances, and distances give a metric for proximity between utterances within a vector space. However this is where the similarities end. LSA makes a significant technological contribution by introducing elements of semantic reasoning into the model. KWI on the other hand derives its measures based on direct keyword overlap.

To find the key words, we compared 67 Experiment 1 transcripts (197,769 words) to a publicly available reference corpus of American business discourse (42,724 words) using the freely distributed WordSmith Tools program in order to identify words that were used with unusual frequency in our UAV task. We then computed the relative frequencies of key words in the 67 Experiment 1 transcripts. From this we determined a set of super key words (words that are prototypical forms of other key words; e.g., "waypoint" is the super key word of "LVN", "H-area", etc.). To further define the key word space, we associated weights to the various super key words (super key word weights are also assigned to all key words in their set). The weights, w(f), were computed from the relative frequency (f) of the super key words relative to one another (borrowing from Shannon's entropy formula: $w(f) = -f \log(f)$). This weighting insured that the less frequent keywords were weighted heavier proportional to their degree of uncertainty than

more frequent key words (e.g., waypoint names were weighted heavier than "restriction"). The key word space consisted of 31 super key words and their weights (see Table 13):

Table 13
Key words and weights

S Key Word	Weight	S Key Word	Weight	S Key Word	Weight
accept	0.01	go	0.06	plo	0.05
after	0.03	good	0.03	radius	0.03
air	0.03	have	0.06	restriction	0.03
altitude	0.05	I	0.08	right	0.03
at	0.03	intel	0.03	roger	0.05
avo	0.05	mile	0.03	speed	0.05
be	0.06	need	0.03	target	0.06
can	0.04	next	0.04	we	0.07
dempc	0.04	now	0.03	waypoints (all)	0.10
effective	0.02	okay	0.06		
get	0.03	photograph	0.06		

Based on KWI vector lengths and word counts for our set of 67 original transcripts, we assessed four relatively low level content measures: Vector – mean vector length of utterances in a transcript; Words – mean number of words per utterance in a transcript; KWIDensity – the ratio of mean vector length to mean word length in a transcript; and Weighted words – the product of mean word length and mean vector length in a transcript. We then looked to see if there were main effects of team or mission on the various KWI vector/word length measures:

Table 14
Effects of Team or Mission on Keyword Indices

DV	Team	Mission
Vector	$F(10,50) = 6.42^{**}$	$F(6,50) = 5.44^{**}$
Words	$F(10,50) = 13.48^{**}$	$F(6,50) = 1.39$
KWIDensity	$F(10,50) = 22.18^{**}$	$F(6,50) = 5.40^{**}$
Weighted	$F(10,50) = 8.55^{**}$	$F(6,50) = 1.93^{*}$

$^{*}p \leq .10$, $^{**}p \leq .01$

Based on the results in Table 14, each of the KWI measures varied significantly over teams and missions. The word count statistic did not vary significantly over missions.

6.3.2.3 Web-based LSA interface. During the course of preparing the LSA tools for analyzing the new data we developed a web interface http://bluff.nmsu.edu/~ahmed/ for automatic analysis of discourse. The interface provides two facilities:

- Discourse Analysis:
  The user can paste or upload discourse (e.g. a team-at-mission transcript) to a web page that automatically returns a new web page with some statistics about the discourse, including LSA coherence and vector length, and suggested tags. In addition, we have augmented the LSA coherence computations that the system can handle to allow for lag n. For example, the original program computed coherence between turn i and turn i+1, now we can compute the coherence between turn i and turn i+n, where n is set by the user. This enables the user to get a fuller picture of the coherence of the team's discourse.

- Predict Performance Scores:
  Programs to predict overall team performance scores based on whole team-at-mission transcripts have been incorporated into our web-based system. A user can now upload or paste in a transcript and the system will automatically return a predicted performance score for the transcript. The user has the choice to choose the number of closest transcripts to be used in the prediction. The number that could generate the best prediction is still to be found. That number will be made as the default value unless the user chooses different number.

As part of this project, a web-based demonstration system was developed that could take incoming transcripts of teams and generated automated performance scores. A screen shot of the system is shown in Figure 3. It illustrates the output of the analysis of a transcript displaying a number of LSA and other statistics that can be useful for characterizing the quality of the team's performance. In addition to basis statistics about the transcript as a whole, it computes the frequencies of the predicted tags. In the discourse section, the predicted tags, their certainty, coherence with the next turn, and vector length (measure of information content of the turn) are shown next to the discourse.

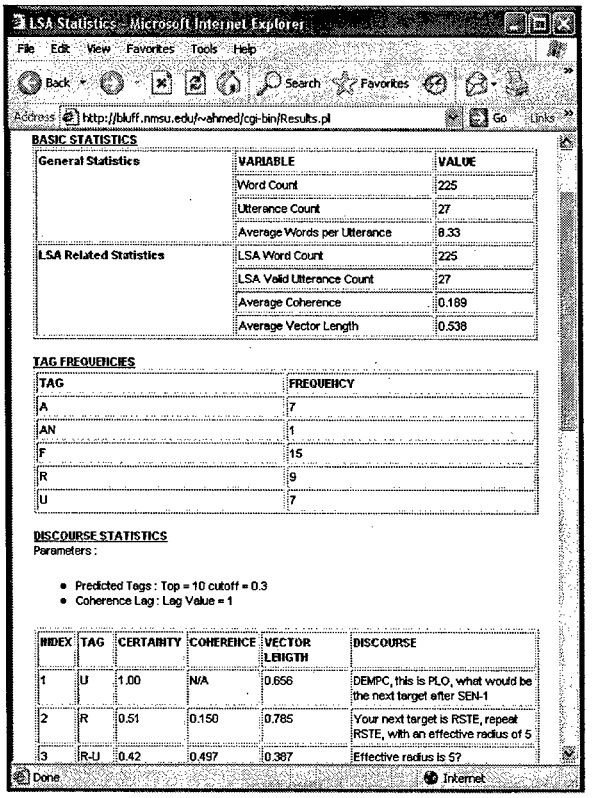


Figure 3. Screen Shot of Web-based System

6.4 Task 2: Examine Method Validity

6.4.1 Method

The primary question associated with this task is whether the content and flow analytic methods generate communication patterns that are predictive of team performance. To answer this question we took the metrics resulting from the application of the communication methods applied in Task 1 and used them to predict (correlations or regression) team performance based on our composite scoring procedure. We expected that if our measures were adequately capturing team communication, and if this communication reflected the team cognition underlying team performance, then the performance and communication measures should be correlated and performance predicted by communications (H2.1; Table 15). Further we expected that results from other communication studies should be replicated using these new metrics (H2.2).

Table 15
Hypotheses Associated with Task 2

Task/Hypothesis Number	Hypothesis	Supported?
2.1	Communication metrics, if valid, should be correlated with performance scores.	Yes
2.2	Functions relating performance and communication metrics should generalize across studies.	Yes

6.4.2 Results

Results are first described for LSA-based measures and then for flow-based (i.e., FAUCET) measures.

6.4.2.1 LSA-Based performance score. As a reminder, to compute the estimated score for t, we take the average of the performance scores of the 10 closest transcripts in the space, weighted by cosines. The holdout procedure was used in which the score for a team's transcript was predicted based on the transcripts and scores of all other teams (i.e. a team's score was only predicted by the similarity to other teams). Our results indicated that for Experiment 1, the LSA estimated performance scores correlated strongly with the actual team performance scores ($r = 0.76$, $p < 0.01$), as shown in Figure 4. Thus, the results indicate that we can accurately predict the overall performance of the team (i.e. how well they fly and complete their mission) just based on an analysis of their transcript from the mission.

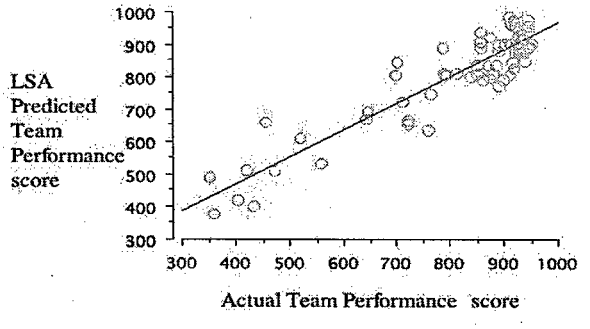


Figure 4. Correlation: Predicted and Actual Team Performance for Experiment 1

To test our algorithm to predict team performance using LSA on whole transcripts, we repeated our experiment using the Experiment 3-Corpus and Experiment 3 semantic spaces. Using the 10 closest transcripts, as before, the LSA estimated scores strongly correlated with the actual scores ($r = 0.75$).

To demonstrate the generalization of our algorithm over varying semantic spaces, we also compared the correlation of estimated and predicted team scores for Experiment 1 and Experiment 3 transcripts using the Experiment 1-3-4 semantic space. The results, shown in Table 16, confirm that performance is not significantly changed by using a larger, more general, semantic space. It further shows that LSA is robust over a range of different sized corpora.

Table 16
Predicted-Actual Scores Correlations When Varying Semantic Spaces

	EXP1	EXP3	EXP1_3_4	Difference
EXP1	0.76		0.77	+1%
EXP3		0.75	0.72	-4%

Generalization of team performance scores for different corpora. While the results were successful for the Experiment 1 corpus, it is important to determine if similar results can be found for the other two corpora. In addition, it is important to determine if the algorithm to predict team performance using whole transcripts can operate successfully by training the algorithm on the performance scores of one corpus in order to predict performance scores on another corpus. This approach would be equivalent to having collected N transcripts from teams flying UAVs on a set of particular missions and then trying to predict a new set of teams performing a different set of missions. Thus, the generalization test, determines how robust such a system could be in more realistic contexts where different teams may have to fly entirely novel missions.

We tested the generalization for the Experiment 3 set of transcripts, by training our algorithm on the performance scores of Experiment 3 performance scores and predicting the performance scores from the other experiment (Experiment 4). Using the 10 closest transcripts, as before, the LSA estimated scores strongly correlated with the actual scores or Experiment 3, showing only a four percent degradation in performance (see Table 17.). Thus, there was a high level of generalization from one training corpus to predicting the performance scores of another.

Table 17
Predicted-Actual Score Correlations When Varying the Training Set

	Training Set		
	EXP 3	EXP 4	Difference
EXP3	0.72	0.66	-4%

Varying the dimension of the semantic space for performance prediction using whole transcripts. In all previous results reported for team score prediction using whole transcripts, the dimension of the semantic space created by LSA was set at approximately 300, which has been shown in to be the best size to capture the complexity of human language on a variety of other tasks. To verify that this holds true for out UAV data, we tested dimensions 100, 200, 300, 400, 500 and 600 on the Experiment 1 and Experiment 3 data sets. The results are shown in Table 18.

Table 18
Correlations between predicted and actual scores as dimension of semantic space varies

Experiment	Dimension	Correlation
EXP 1	100	0.786803723
EXP 1	200	0.760132232
EXP 1	**300**	**0.786847128**
EXP 1	400	0.745200698
EXP 1	500	0.743739265
EXP 1	600	0.748404337
EXP 3	100	0.113699842
EXP 3	200	0.628493896
EXP 3	**300**	**0.738088108**
EXP 3	400	0.708010772
EXP 3	500	0.722278549
EXP 3	600	0.735669498

These preliminary results appear to confirm that the best choice for the dimension of the LSA- created semantic space is 300.

Automatic speech recognition. We showed that LSA works adequately using transcripts based on current speech recognition software, despite the limitations of the state-of-the-art in speech recognition. We used speech recognition error rates from Schmidt-Nielsen, Marsh, Tardelli, Gatewood, Kreamer, Tremain, Cieri, Strassel, Martey, Graff, and Tofan (2001).We then introduced similar errors in existing (typed) transcripts to represent synthetic speech errors (e.g., insertions, deletions, substitutions). Finally, we tested LSA's effectiveness at predicting team scores at different error rates. Results vary with transcription error rate, and show 80% reliability at speech recognition error rates up to 57%.

6.4.2.2 LSA-Based automatic tagging. In order to test the ability of our automatic tagging algorithm to generalize, we trained a new annotator. He was trained on the Experiment 1 corpus and in testing achieved good agreement with the previous annotators: Kappa was 0.72. Given this level of agreement we had him tag 20 randomly selected transcripts from each of Experiment 3 and Experiment 4 (approximately 24% of

the total discourse in these corpora). We were then able to compare our automatically predicted tags for Experiment 3 and Experiment 4 to his tags (see Table 19).

Table 19
LSA+ - Annotator Agreement

	EXP 1	EXP 3	EXP 4
Kappa	0.53	0.56	0.54
C-value	0.63	0.66	0.64

The results indicate that humans can consistently use the Bowers tag set across the three corpora and that the LSA+ algorithm can consistently predict the tags.

We were also able to show generalization across semantic spaces: training on the tags in Experiment 1 to predict tags in Experiment 1, produced equivalent Kappas (to two decimal places) using the Experiment 1 and Experiment F1-3-4 semantic spaces. In addition we varied the set of tags used for training. In the Experiment 1-3-4 semantics space, predicting tags for the Experiment 3 corpus showed only a 5% degradation in performance when the system was trained on the Experiment 1 tags rather than on the Experiment 3 tags (Table 20). We believe this demonstrates the robustness and ability to generalize, at least within the UAV-STE domain, of the LSA+ algorithm.

Table 20
Generalization of tagging based on other semantic spaces

	Training Set		
	EXP 3	EXP 4	Difference
EXP3	0.72	0.66	-4%

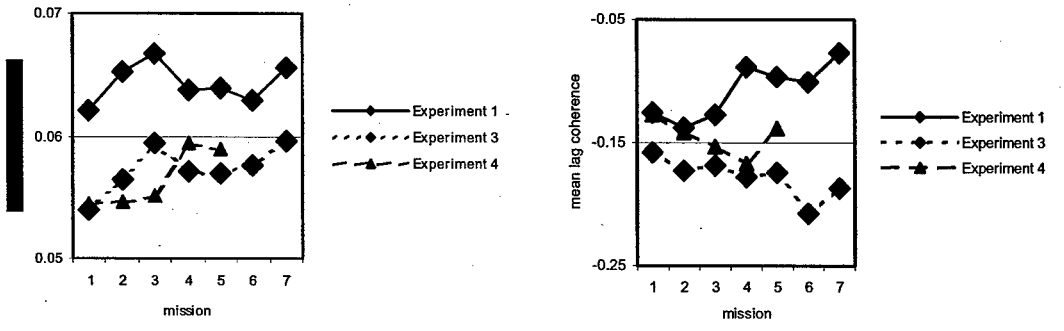
6.4.2.3 LSA-Based density and lag coherence. As described previously, the transcript density measure is the ratio of average LSA vector length to average number of words per utterance per mission. Transcript density is thus a relative measure of the average domain relevant content of utterances within a mission transcript.

The lag coherence measure is the least-squares slope between average LSA cosine and lag between utterances up to 35 utterances away. Thus this measure is an average correlation over a 36-utterance moving window taken for each transcript.

The mean transcript densities and lag coherence by experiment are presented in Table 21. Figure 5 shows mean transcript density (a) and mean lag coherence (b) by mission for Experiments 1, 3, and 4.

Table 21
Mean Transcript Density and Lag Coherence for Experiments 3 and 4

	Experiment 3		Experiment 4	
	Density	Lag Coher	Density	Lag Coher
N	85	85	60	85
Mean	0.057	-0.179	0.057	-0.146
SD	0.005	0.066	0.005	0.045



(a) (b)

Figure 5. Mean transcript density (a) and mean lag coherence (b) by mission for Experiments 1, 3, and 4.

For content metrics we examined two types of validity. First, we examined whether or not the functional forms (i.e., number and direction of terms) were consistent across experiments, and second we examined predictive validity. Functional form was analyzed using baseline results from Experiment 1. Mean squared prediction error (MSPE) across models was examined by using estimates from the baseline Experiment 1 transcript density and lag coherence models to predict Experiment 3 and 4 team performance (outcome) scores. The variance of the residual score, [(performance – predicted performance)^2]/number of observations, was the estimate of mean squared prediction error. The average ratio of MSPE/MSE, where MSE is taken from the original model, is reported. These results are presented in Table 22.

Table 22
Validation Results for Transcript Density and Lag Coherence

Task 2

Measure	Experiment	Functional Form	MSPE
Density	1	pos linear, neg quadratic	1.5
	3	neg linear, pos quadratic	
	4	neg linear, pos quadratic	
Lag Coher	1	pos linear	2.5
	3	pos linear	
	4	neg quadratic	

Overall, these results suggest that the density measure exhibited higher validity than the coherence measure. First, while inconsistent with Experiment 1, the functional forms were similar for the density measure between Experiments 3 and 4. Lag coherence had a different pattern with Experiments 1 and 3 leading to similar functional forms, which were different from Experiment 4. These results provide partial support for H2.2. The overall prediction variance however, was on average smaller for the density measure.

In addition to these analyses, we also asked ourselves the question what is the relation between the basic LSA density component, vector lengths, and "leaner" measures of content such as word counts and keyword counts, and how might this relate to the overall validity of the density measure. Correlations over 20,545 utterances between these measures are shown in Table 23. The results indicate that it is highly likely that LSA vector length derives most of its variance from the length, in number of words, of utterances. This could result from one of two things. First, LSA vector *is* a measure of word counts. Or second, and most probably, task-related talk comprises the lion's share of the utterances in our transcripts. In either case however, this result suggests that the density measure may not have as good "face" validity as we initially thought, since it is apparently comprised of the ratio of two measures of word counts, however they are derived. On the other hand, a measure correlated with LSA vector lengths, KWI vector lengths, was found to share less variance with straight word counts. In a subsequent set of analyses we report on using KWI vector lengths in density ratios, in comparison to LSA density ratios.

Table 23
Correlations Between LSA Density Component and Other Content Metrics

	LSA Vector Length	KWI Vector Length
Word Count	.944*	.699*
LSA Vector Length	-	.725*

Note. $N = 20{,}545$; * $p < .001$

6.4.2.4 FAUCET metrics. The FAUCET methods of most value were the measures derived from our Dominance method, CRP, ProNet, and CHUMS. We also analyzed the Process Surrogate method for the first study, but dropped it in subsequent analysis. We decided that the measure was too similar to dominance and CRP. Overall the flow-based metrics were predictive of team performance.

The way we handled FAUCET comparison to basic measures was through covariance. The FAUCET findings used total amount of speech as a covariate in the analyses, so that any FAUCET findings are known to be above and beyond the relationship between the basic flow measure, and the criterion.

Specific comparisons between experiments for the FAUCET measures (H2.2) are found in the performance section. In general, the hypothesis was supported with most of the measures. For example, CHUMS measures showed that number of models (a measure of communication instability) was negatively correlated with performance in both of the latter studies. This was also found for the ProNet stability measure, which was positively related to performance in both of the latter studies.

Dominance findings. Only T-1 dominance scores are independent, where T is the number of team members. The last score is determined by the others. The set of T-1 independent dominance scores are adequate predictors of performance. This includes analysis of these data in four studies. Statistically detectable mission-by-mission correlations ranged from $R = .63$, ($F(2, 15) = 4.90$, $p=.023$) to $R = .75$, ($F(2, 8) = 5.18$, $p=.036$). More specific analyses for individual dominance scores revealed that AVO dominance tended to be a good predictor of performance, though valence was inconsistent (e.g. for Mission 4 of the third study, $R = .441$, $F(14) = 3.37$, $p = .088$; for Mission 3 of the fourth study, $R = -.415$, $F(1, 15) = 3.12$, $p = .098$).

CRP findings. The set of all three CRP scores are adequate predictors of performance. This includes analysis of these data in four studies. Statistically detectable mission-by-mission correlations ranged from $R = .65$, ($F(3, 14) = 3.14$, $p=.062$) to $R = .81$ ($F(3, 7) = 4.61$, $p=.044$). More specific analyses indicated that PLO's CRP score was the most important predictor (e.g. for the omnibus test of the third study, after adjusting for repeated measures, $R = .20$, $F(1, 90) = 3.946$, $p = .050$; PLO's information component, $R = .22$, $F(1, 87)=4.421$, $p = .038$).

ProNet findings. The minimum, median, sum, and maximum chain lengths for each team-at-mission were the best predictors of performance. Statistically detectable mission-by-mission multiple correlations ranged from .52 ($F(1, 9) = 3.42$, $p = .098$) to .79 ($F(2, 8) = 5.07$, $p = .051$). For the first study, a model including minimum, median, and sum tended to yield better predictions for early missions, while the minimum alone was the best predictor for later missions. For the third study, the sum and maximum tended to be good predictors throughout. Surprisingly, very few good predictions were found for the fourth study. Analyses of communication flow sequence using ProNet detection have revealed that some new speech sequences that are predictive of performance. Results

indicate that longer detectable sequences (and hence more regular communication patterns) are generally linked to better performance and process.

CHUMS findings. Number of clusters and clusters per minute are adequate predictors of performance. Statistically detectable mission-by-mission correlations for number of clusters ranged from $R = .54$ ($F(1, 9) = 3.63$, $p = .089$) to $R = .61$ ($F(1, 8) = 4.75$, $p = .061$), or $R = .45$, ($F(1, 16) = 4.10$, $p = .060$) to $R = .49$, ($F(1,8) = 2.56$, $p = .148$) for clusters-per-minute. Number of clusters and clusters per minute are also adequate predictors of situation awareness. Statistically detectable mission-by-mission correlations were $R = .65$, ($F(1, 8) = 5.93$, $p = .041$), for clusters-per-minute, and $R = .62$, ($F(1, 8) = 4.982$, $p = .056$), for number of clusters.

Process surrogate findings. All 10 scores were correlated with the team performance score. The analysis was repeated in a replication study. Correlations ranged from -.001 to .95 for the dominance-based aggregates, .01 to .79 for the CRP-based aggregates, -.04 to .75 for the observational process-based aggregate, and .50 to .93 for the simple arithmetic mean. Most predictors were approximately as good as the baseline measure. Those that were better predictors than the baseline were the arithmetic mean of Dominance during various missions (r's from .56 to .95), and the minimum CRP score during initial missions (r's from .54 to .79).1)

6.4.3 Conclusions

- Our communication analysis metrics correspond to team effectiveness supporting H2.1.
- LSA-based performance scores correlate with actual team performance (r=.75) supporting H2.1.
- LSA-based performance scores generalize over different semantic spaces and training sets within the same task domain
- LSA-based performance scores are robust to errors due to speech recognition software
- LSA can be used to consistently predict tags (i.e., content codes).
- LSA-based tagging generalizes across semantic spaces and training spaces
- LSA-based density and coherence functions replicate across two of the three experiments lending partial support to H2.2.
- Communication stability predicts performance, using CHUMS and ProNet measures, across studies, supporting H2.2.
- LSA-based density exhibits higher validity that LSA-based lag coherence
- LSA vector length is highly correlated with the leaner measure of word count.
- ProNet results indicate that longer detectable sequences (and hence more regular communication patterns) are generally linked to better performance and process.

6.5 Task 3: Examine the Team Performance-Communication Relationship

6.5.1 Method

This task is related to the second task, but we will be taking a closer look at the nature of the best-fitting function relating performance to each communication metric, and/or any combination of individual metrics. Do these patterns make sense in regard to previous studies and effective or ineffective teams? Taken together, does the set of content and flow metrics tell a coherent story about the relationship between team performance and team communication? This task will involve additional regression analyses and extensive interpretation. Additionally, multiple communications metrics may be structured as a multivariate fixed variable and used as a conglomerate predictor of performance. This endeavor serves the very specific purpose of detailing exactly how much performance variance can be accounted for by taking each of our measures while minimizing the number of parameters estimated in the model. This endeavor should lead to more stable performance prediction across samples.

Although, as noted previously many of our communication metrics are new and have not been applied in previous team research. However, based on some related results in the team communication literature, we can begin to formulate some hypotheses regarding the specific relations between communication and team performance. We predict that teams with higher performance scores will have greater disparity among communication flow dominance scores because they will tend to have a leader who knows the task (H3.1). In addition we predict that high-performing teams will have (according to the flow sequence analysis) more completed speech cycles (i.e. Person A begins, Person A ends, Person B begins, person B ends) and fewer interruptions than low-scoring teams (H3.2). We also expect high-performing teams to have longer mean chain lengths in the sequential flow analysis than low-performing teams because they understand the task well enough to have a rehearsed routine (H3.3). In terms of flow stability, we predict that high-performing teams will have fewer patterns of communication frequency due to their superior teamwork knowledge and skills (H3.4). We also predict content differences in the communications of good vs. poor teams. Based on the literature, better teams will follow assertions or action statements with acknowledgements as based on LSA coding, more than ineffective teams (H3.5). Further, based on preliminary analysis of data from a different study, better teams in terms of performance will have communication efficiency scores that fall within a mid-level range (H3.8). As team skill is acquired we predict that individual teams will start out as inefficient and then proceed to being overly "efficient" and finally reach the point of optimal efficiency (H3.7). In terms of LSA-based coherence, effective teams will demonstrate greater coherence at longer lags, suggesting decreased topic shifting during missions compared to ineffective teams who will demonstrate less coherence at longer lags, suggesting a great deal of topic shifting over a mission (H3.8). Individually, teams should start out with small lag coherence and at asymptotic levels of performance (mission 4-5) should demonstrate greater lag coherence (H3.9).

Table 24
Hypotheses Associated with Task 3

Task/ Hypothesis Number	Hypothesis	Supported?
3.1	Teams with higher performance scores will have greater disparity among communication flow dominance scores.	No
3.2	High-performing teams will have (according to the flow sequence analysis) more completed speech cycles and fewer interruptions than low-scoring teams.	No
3.3	High-performing teams should have longer mean chain lengths in the sequential flow analysis than low-performing teams.	Yes
3.4	High-performing teams will have fewer patterns of communication frequency (in terms of flow stability metric-CHUMS) than low performing teams.	Yes
3.5	Better teams will follow assertions or action statements with acknowledgements as based on LSA coding, more than ineffective teams.	Yes
3.6	Better teams in terms of performance will have communication efficiency (density) scores that fall within a mid-level range.	No
3.7	As team skill is acquired, we predict that individual teams will start out as inefficient (low density) and then proceed to being overly "efficient" (high density) and finally reach the point of optimal efficiency (mid-level density).	Yes
3.8	Effective teams will demonstrate greater coherence at longer lags compared to ineffective teams who will demonstrate less coherence at longer lags.	Yes
3.9	Individually, teams should start out with small lag coherence and at asymptotic levels of performance, should demonstrate greater lag coherence.	Yes

6.5.2 Results

6.5.2.1 Team Performance Prediction Using Automatically Generated Discourse Tags. The results LSA-based tagging developed in Task 1 showed that the methods could provide accurate characterization of the type of utterances made by team members. In Task 2 we demonstrated the generality of these methods over different training and

semantic spaces within the same domain. In order to look at the relationship between team performance and communication, it is necessary to determine if types and frequencies of utterances made by team members are indicative of performance. We computed correlations between the team performance score and tag frequencies in each team-at-mission transcript in the Experiment 1 corpus.

The tags for all 20,545 utterances in the Experiment 1 corpus were first generated using the LSA+ method. The tag frequencies for each team-at-mission transcript were then computed by counting the number of times each individual tag appeared in the transcript and dividing by the total number of individual tags occurring in the transcript. The results of the single tag-performance score correlations are shown below (Table 25). We were able to identify several LSA predicted codes that correlate with team performance.

Table 25
Correlation of single discourse tags to team performance

SINGLE TAG	PEARSON CORRELATION	SIG. 2-TAILED
Acknowledgement	0.335	0.006
Fact	0.320	0.008
Response	-0.321	0.008
Uncertainty	-0.460	0.000

Table 25shows that the automated tagging provides useful results that can be interpreted in terms of team processes. Teams that tend to state more facts and acknowledge other team members more tend to perform better. Those that express more uncertainty and need to make more responses to each other tend to perform worse. These results are consistent with those found in Bowers et al. (1998), but were generated automatically rather than by the hand-coding done by Bowers.

Using the methodology discussed above, we also computed tag bigrams (adjacent two-tag sequences) and correlated them with team performance. The significant results are shown below in Table 26

Table 26
Correlation of discourse tag bigrams to team performance

TWO-TAG SEQUENCES	PEARSON CORRELATION	SIG. 2-TAILED
Acknowledgement - Fact	0.263	0.031
Fact - Acknowledgement	0.259	0.034
Uncertainty - Response	-0.270	0.027
Uncertainty - Uncertainty	-0.414	0.000

Table 26 shows that tag bigrams also provide useful results and insights into team processes. Most striking is the correlation between teams where an expression of uncertainty is more frequently followed by another expression of uncertainty with poor overall team performance. Sequences of factual and acknowledgement seem to contribute to improved team performance. We hypothesize that the contribution of the uncertainty-response sequence to poor team performance is directly related to the increased frequency of single uncertainty statements, rather than the combination of the tags, since we know increased uncertainty correlates with poor performance and given a statement expressing uncertainty, it is quite likely that the next statement will be a response.

6.5.2.2 Relationship between the keyword method and performance. The KWI measures described under Task 1 were examined for correlation with team performance using data from the earlier Experiment 1. Correlations between mean word length, and KWI transcript density, as well as weighted words and team performance were found to be significant as shown in Table 27.

Table 27
Keyword indices and performance correlations

	KWIDensity	Weighted	Vector	Words
Performance	.209*	-.217*	-.086	-.297**

N = 67, * $p \leq .10$, ** $p \leq .05$

As part of these analyses we also re-fit the LSA transcript density team performance regression models from Jamie Gorman's MA thesis (Table 28) using the same Experiment 1 data set. In the thesis, mean vector lengths and mean word lengths were combined as either a ratio (i.e., density) or a product (i.e., weighted words), and this was used as a covariate along with team and mission effects to predict performance (Table 29).

Table 28
Regression results from Jamie Gorman's thesis – Using LSA density to predict team performance outcome

Performance – LSA Density Relationship		
	LSA Density	LSA Density^2
Performance	t (47) = .477	t (47) = -2.49*

* $p \leq .05$

Table 29
Regression results of using KWI metrics to predict team performance

Performance – KWI Relationship		
	KWIDensity	Weighted Words
Performance	t (48) = 1.62	t (59) = -2.97*

* $p \leq .01$

The KWI density score was a marginal predictor of team performance rate ($p = .11$). As in the thesis this score related to performance best as a quadratic trend, however this time as a positive quadratic term. The weighted words score did a good job of predicting team performance (the linear trend). However these estimates were only significant when the effects of team membership on performance were left unaccounted for. This might be indicative that either "team" and KWI weighted words are collinear in terms of performance, or "team" moderates the relationship between KWI weighted words and team performance (i.e., "team" is a good predictor of performance with KWI in the model). The former might be tentatively more appealing since these analyses were correlational in nature. In any case we were encouraged with these results indicating that KWI content analysis may play some role in explaining UAV STE team performance.

6.5.2.3 Relationship among transcript density, lag coherence, and team performance. In order to estimate the relationship between team performance and team communication operationalized as transcript density and lag coherence, linear regression was used. Specifically, each of the two communication measures was used as an independent variable to predict team performance at each mission. Table 30 gives the results of these analyses.

Table 30
Performance – Communication Relationships: Transcript Density and Lag Coherence

Performance -- Density/Lag Coher Relationship		
DV	Experiment 3	
	Density	Lag Coher
Performance	t (82) = 1.977*	t (82) = -.074
	Experiment 4	
	Density	Lag Coher
Performance	t (58) = .510	t (83) = -3.390**

Note. * $p < .10$; ** $p < .01$

Across Experiments 3 and 4, higher density predicted higher team performance, contrary to H3.6, although this finding only reached significance in Experiment 3. Ostensibly this means that more "UAV-talk" relative to total talk is associated with higher levels of team performance, however there may be difficulties with this interpretation given the relationship between word counts and the LSA vector lengths that comprise this measure;

specifically that density is the ratio of two word count measures that were measured on different scales (please refer to Task 2 for more information).

The effect of Mission on transcript density was tested in order to address H3.7. For Experiment 3, the (linear) effect of Mission on transcript density was significant (F (6,78) = 1.831, p = .104). For Experiment 4, the (linear) effect was also significant (F (4,55) = 2.881, p < .05). Figure 6 depicts the mean results over missions. As predicted, transcript density started low, increased dramatically, and there is some evidence (i.e., Missions 4-6 for Experiment 3, Mission 5 for Experiment 4) that teams returned toward a more moderate level of transcript density after an initial, dramatic increase. However, it should be noted that changes subsequent to Mission 4 are confounded with increased workload (Task 4).

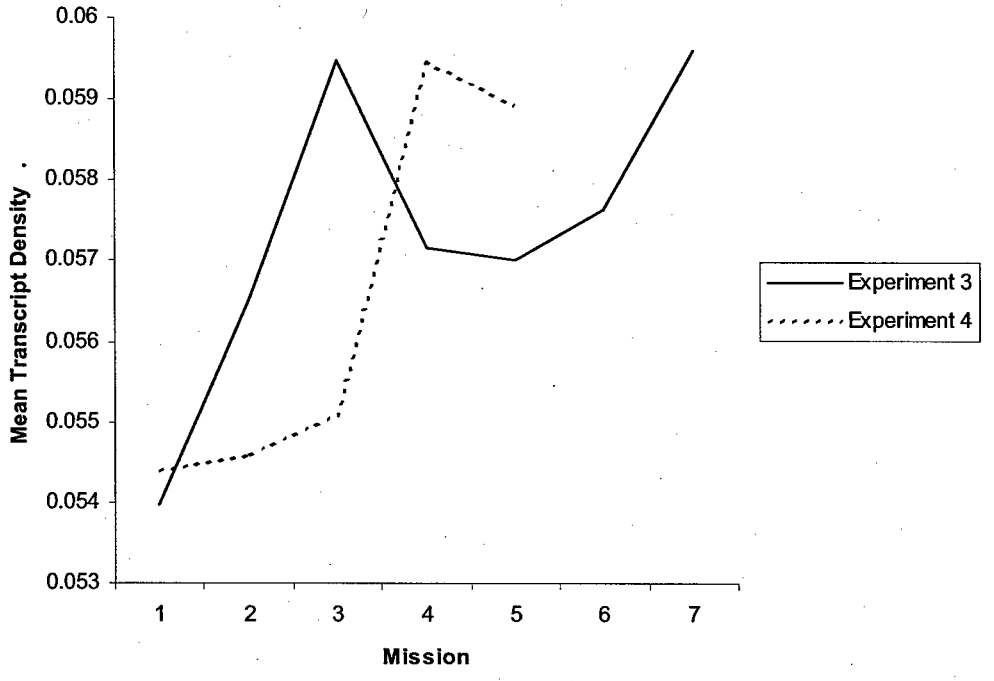


Figure 6. Mean transcript density as a function of mission in Experiments 3 and 4.

For lag coherence, the performance relationship was also consistent across Experiments 3 and 4, although this finding only reached significance in Experiment 4. (Please note that this is inconsistent with the findings from Task 2. In general usage there would be no reason to assume these relationships are polynomial.) Specifically, the negative relationship suggests that as lag coherence becomes increasingly negatively sloped, team performance improves. In laymen's terms, as conversations become related over time, we observe increases in team performance, supporting H3.8.

The effect of Mission on lag coherence was tested in order to address H3.9. For Experiment 3, this effect was not significant (F (6,78) = .722). For Experiment 4, the relationship was significant (F (4, 80) = 1.988, p = .104). Figure 7 shows the mean level of lag coherence across missions for Experiments 3 and 4. It appears that the

developmental trend predicted in H3.9 was exhibited in the low workload missions (1-4), but was interrupted by the high workload manipulation (Mission 5).

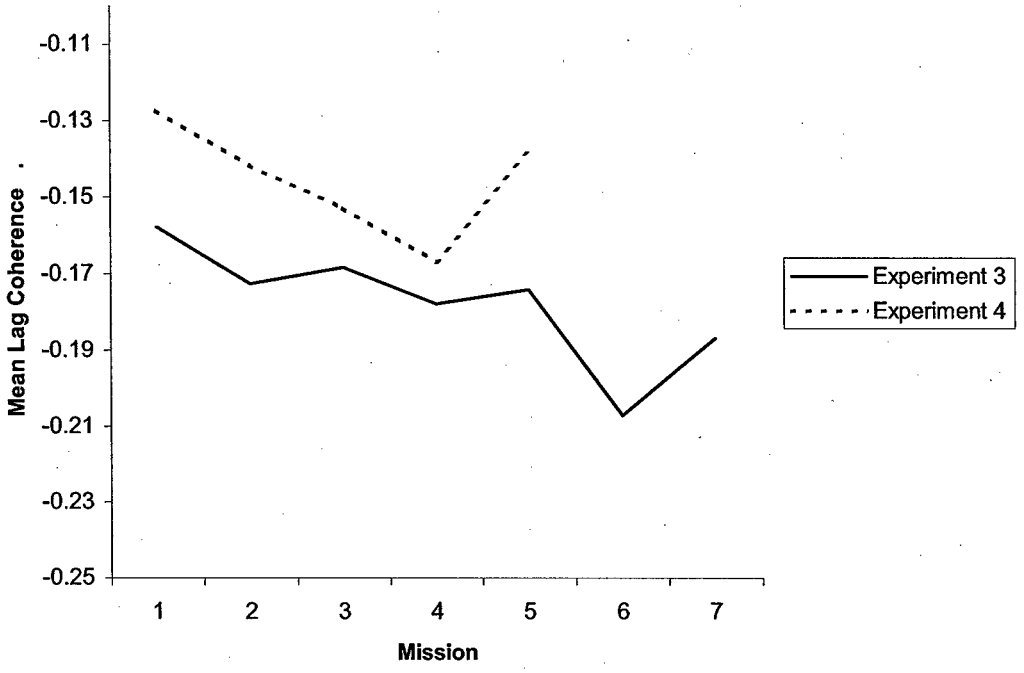


Figure 7. Mean lag coherence as a function of mission in Experiments 3 and 4.

For both Experiments 3 and 4, the fact that one but not the other of these two content metrics was significantly related to team performance is interesting. It can be speculated that this may be due in part to the longer duration of high workload in Experiment 3 versus Experiment 4. In particular, the number of high workload observations could impact the team performance – communication relationships because team performance decreased under high workload. Paradoxically then, if there are significant workload effects of these metrics (please refer to Task 5) then we should expect an attenuation of the team performance – communication relationship for Experiment 4, given the relatively small number of high workload observations.

6.5.2.4 FAUCET Metrics. *CHUMS findings.* Team performance effects were detectable by CHUMS model counts in mission-by-mission tests for both Experiments 3 and 4 at performance asymptote Mission 4 (Exp. 3: $F(1,14) = 5.81, \beta = -.541, p = .03, R^2 = .293$; Exp. 4: $\beta = -.422, p = .09, R^2 = .178$). In Experiment 4 a similar effect was found at Mission 3 (Exp. 3: $F(1, 15) = 4.830, \beta = -.494, p = .04, R^2 = .244$). For team process behaviors at critical events (e.g., communicating the correct information in the correct order) omnibus (i.e., over Experiments 3 & 4) negative correlations were found between model counts and team process, with Missions 3 ($r = -.46$) and 4 ($r = -.44$) being the highest. CHUMS model counts is one of our communication consistency measures. Hence, this finding indicates that teams with more consistent communication, also tend to have better team process supporting H3.4.

Procedural networks (ProNet). Analyses of communication flow sequence using ProNet detection has revealed that some specific speech sequences are predictive of performance. For instance, ProNet detection of PLO-DEMPC cycles (the Pbegin-Pend-Dbegin-Dend sequence of speech events) were found to negatively predict performance at an omnibus level for Experiment 3 (F (1, 94) = 3.692, p = .058, M_{absent} = 469.44, $M_{present}$ = 442.45). This effect was smaller, but in the required direction for Experiment 4 (F (1, 64) = 2.197, p = .143, M_{absent} = 544.10, $M_{present}$ = 523.22). This implies that complete speech sequences beginning with PLO and ending with DEMPC are predictive of poor performance. This contradicts H3.2. Since these were the only complete cycle and/or interruption results that were found, we conclude that H3.2 was not supported.

Other predictive results for the ProNet measures indicate that longer detectable sequences (and hence more regular communication patterns) are generally linked to better performance and process (supporting H3.4).

Flow patterns and glitch adaptations. During the mission in which the glitch was introduced, teams exhibited communication behaviors consistent with adaptation. Most notable were the ProNet measures. The DEMPC-to-AVO channel cut was associated with more DEMPC-to-PLO complete utterance cycles, more PLO-to-AVO utterance cycles, and fewer DEMPC-to-AVO cycles. Other communication findings were also consistent with adaptation, such as an increase in total communication patterns (CHUMS measures), PLO communication quantities deviating from the pre-determined norms (CRP measures), and a shift in the Dominance measures toward DEMPC dominance.

Dominance, CRP, and process surrogate. No outstanding findings were made for Dominance, CRP, or Process Surrogate. The few findings that were revealed, tended to be at a mission-by-mission basis, and were not replicated across studies. Hypothesis 3.1 was that the dominance measure should be predictive of performance. Since it was not, H3.1 is rejected.

The process surrogate measure was rejected from further analysis, because it is conceptually too similar to the dominance and CRP measures.

6.5.2.5 Combined methods. One approach to combining the methods was to incorporate all of them, or a representative from each method, into a single regression equation. This way, the variance shared among the predictors can be removed, and we can assess their unique contribution to performance. For the third study, the following model resulted from a procedure that was based partly on automatic variable selection, and partly on an attempt to maximize diversity in represented predictors. No suppressor variables were required. The result is displayed Table 31.

Table 31
Regression results from combined FAUCET predictors

$R^2 = .24, F(3, 90) = 9.468, p < .001.$

Process Surrogate:
The Arithmetic mean of Dominance weighted individual performance (AritD)
$B = 991.346, \beta = .224, t(90) = 3.981, p < .001$
CRP:
The extent to which AVO did not receive excess passed communication (Achat)
$B = -1351.033, \beta = -.160, t(90) = -2.346, p = .021$
ProNet:
Maximum chain length among the set of identified chains (Max)
$B = -5.189, \beta = -.101, t(90) = -2.638, p = .010.$

We interpret this model as bearing the following implications:

+AritD→Communication dominance of team members predicts their impact on team performance
-Achat→the AVO should speak and/or be spoken to more, rather than less
-Max→ Chains should not be too long or in other words, speech acts should not be too scripted

The last finding is particularly interesting, because the ProNet findings, when taken by themselves, generally suggest that more regularly scripted communication patterns are associated with better team performance and process.

Another combined approach is through Hidden Markov Models. Our Hidden Markov models (HMM) assumed that the flow of discourse content among team members could be modeled using first order Markov transitions, and that this process emits codes from a discrete tagging alphabet. Briefly, HMM's consist of a set of parameters, a state space, and an alphabet. Our parameterization includes the conditional probabilities of each speaker following another (transition matrix **A**), an initial speaker probability vector (**i**), and the conditional probabilities of each code being observed while the process is in each of the states (emission matrix **B**). The state space was constructed by interconnecting all content-emitting sources of a team (see Figure 8). The hidden Markov process was also assumed to be ergodic and stationary. The alphabet of possible signals emitted from this process were U-uncertainty, F-fact, A-acknowledgement, AN-take action, R-response, NT-off task, P-planning, EXP-intelligence, and C-compound, where C is some combination of the other 8 codes.

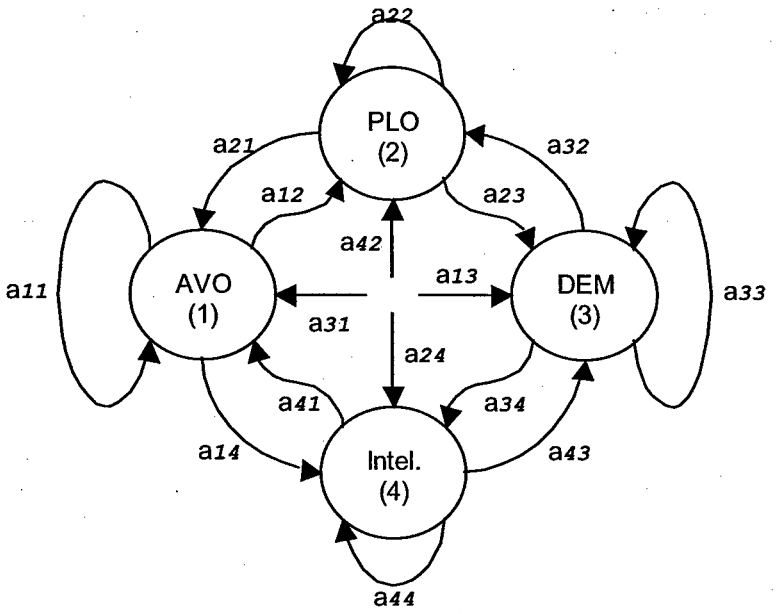


Figure 8. State space diagram with transitions (the **A**'s) for a CERTT team's Markov model.

In order to obtain maximum likelihood (ML) estimates for model parameters, we used the iterative Baum-Welch procedure. Essentially this procedure maximizes the probability of an observed training sequence by iteratively adjusting the models parameters. Baum-Welch however, only guarantees local ML estimates, a significant problem with larger parameter spaces. Thus we also incorporated a "hill-climbing" strategy in which we varied the starting values of the parameters, and iterated until we converged on a set of estimates. Additionally, we started with observed emission probabilities in order to initiate the algorithm in the right directions. In Figure 9, likelihood over iterations is illustrated for three sets of starting parameter values for Experiment 1 Team 2 (5670 coded utterances). Set # 3 is the actual observed transitions among speakers in the data set (included for comparison), while sets 1 and 2 start with **A** values for equal-probable transitions among the team member content sources (.45 and .33 respectively) and near zero transitions for team members following themselves (.01). In both sets 1 and 2 team members following Intelligence (i.e., the experimenter) were equal-probable (.10 and .33 respectively) as is Intelligence following any team member (.10 and .33 respectively). Across the starting sets, ML estimates are obtained within a few iterations (7). The estimates across the hypothetical starting configurations are not in complete agreement (in many cases only to the first or second decimal place), therefore the final estimates for elements of **A** and **i** were averages across solutions (Note that **B** always starts the same and agreement across starting configurations is much higher). Examples of estimates are given in the table below Figure 9.

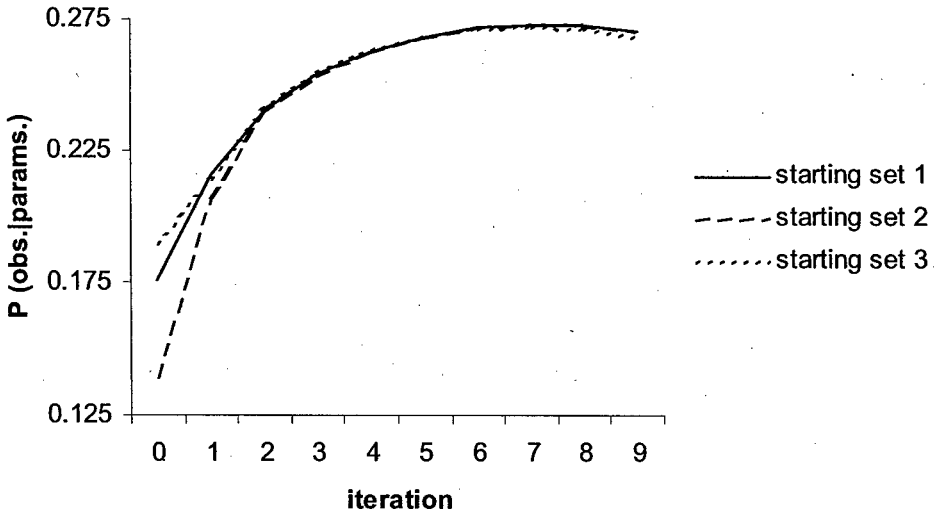


Figure 9. Likelihood of training sequence over Baum-Welch iterations.

A =

	AVO	PLO	DEM	INT
AVO	0.0220282	0.6069003	0.3707547	0.0003168
PLO	0.6829903	0.0160469	0.3002417	0.0007212
DEM	0.5706259	0.4170825	0.0106203	0.0016713
INT	0.0078401	0.0051787	0.0265421	0.9604439

B =

	A	AN	C	EXP	F	NT	P	R	U
AVO	0.2487994	0.0005501	0.4281435	0.0161844	0.2530311	3.347E-15	5.871E-16	0.0000345	0.0532572
PLO	0.2490362	0.0016990	0.3540993	0.0259375	0.2866269	0	0	0.0000438	0.0825573
DEM	0.1888008	0.0018488	0.4392849	0.0365911	0.3021784	7.719E-16	1.015E-15	0.0000828	0.0312131
INT	0.0173992	0.0000638	0.0692706	0.7979959	0.1056288	1.461E-14	0	0.0000117	0.0096300

i =

AVO	PLO	DEM	INT
0.2416041	0.398213	0.3458598	0.0143231

The relationship between a teams' Markov model and its' performance is not clear, nor have extensive efforts yet been invested in addressing this relationship. However one initial effort, in which "high outcome performance" parameter estimates were compared to individual team estimates have indicated that DEMPC – PLO sequences are more likely in the high performance estimates than in a nominal mission transcript. Also the high performance transcript estimates reveal lots of use of acknowledgments, uncertainty, and factual type statements. Further, in poorer performance transcripts we found a high probability for response and uncertainty, but low probability for acknowledgements. These latter analyses have indicated that a good team's uncertainties are followed more often by facts. A poorer team's uncertainties are less likely to be followed by a fact. These results corroborate and further delineate the automatic tagging results described previously in this section.

6.5.3 Conclusions

- Teams that tend to state more facts and acknowledge other team members more tend to perform better. Markov analyses also supports this claim.
- Those that express more uncertainty and need to make more responses to each other tend to perform worse. Markov analyses also supports this claim.
- We were encouraged with these results indicating that KWI content analysis may play some role in explaining UAV-STE team performance
- Higher density (i.e., more "UAV talk") predicted higher team performance
- As predicted, transcript density started low, increased dramatically, and there is some evidence (i.e., Missions 4-6 for Experiment 3, Mission 5 for Experiment 4) that teams returned toward a more moderate level of transcript density after an initial, dramatic increase.
- As conversations become related over time, we observe increases in team performance
- In the low workload missions teams tend to start out with small lag coherence and at asymptotic levels of performance demonstrate greater lag coherence.
- Teams with more consistent communication, also tend to have better team process
- More regular communication patterns are generally linked to better performance and process
- Adapting to a cut communication channel by creating more communication patterns was beneficial for relative performance
- FAUCET methods predict team performance better than simple, low level communication quantity methods.

6.6 Task 4: Investigate Co-located (F2F) vs. Distributed Collaboration

6.6.1 Method

In both Experiments 3 and 4 teams were assigned to either co-located or distributed environments in which they stayed for the 7 or 5 missions, depending on the experiment. Under this task we explore the impact of this environmental factor on communication patterns. The comparison of the increasingly common distributed environment to the face-to-face condition provides a baseline against which to assess the impact of geographic separation. The communication metrics are each examined individually and jointly, for differences due to this manipulation. In cases in which differences exist, the nature of those differences is explored. In addition, within a condition, the relationship between performance and communication is examined. It is possible that communication is more critical for performance in environments that are not face-to-face. In addition, patterns of communication are examined in each condition over time to identify any evidence of adaptation via communication behavior. Specifically, analyses include uni- and multivariate inferential statistical testing, depending on the hypothesis being considered.

Numerous possible hypotheses can be tested in this context. For example, it is possible that communication is more critical for performance in environments that are not face-to-face than those that are and so performance-communication relationships identified under Task 3 should be strongest in the former condition (H4.1). It is also possible that distributed teams will spend more time discussing teamwork and non-task related topics than co-located teams, because they will feel the need to become acquainted with one another (Walther, 1996). This should become apparent in differences in LSA-based code frequencies for co-located vs. distributed conditions (H4.2). Because teams in distributed conditions should gradually acquire the missing teamwork and interpersonal information, the differences found due to mission environment should dissipate with experience (H4.3).

Table 32
Hypotheses Associated with Task 4

Task/Hypothesis Number	Hypothesis	Supported?
4.1	Performance-communication relationships identified under Task 3 should be stronger in the distributed condition than the co-located condition.	No
4.2	We expect LSA-based coding to reflect more teamwork and non-task related communications among distributed teams than co-located teams.	Yes
4.3	The communication differences found in 4.2 should dissipate with experience.	Yes

6.6.2 Results

6.6.2.1 Using LSA to predict whether teams were co-located or distributed. The goal of using LSA in investigating F2F vs. distributed collaboration was to try to model differences between the F2F and distributed groups in order to determine whether there were measurable differences in language between the groups. We were not able to find significant differences using whole-transcript discourse to predict whether teams were co-located or distributed using LSA.

6.6.2.2 Co-located vs. distributed transcript density and lag coherence. The LSA-based density and lag coherence measures were analyzed for effects of the co-located vs. distributed manipulation in Experiments 3 and 4. Table 33 lists the number of observations, means, and standard deviations of each score for each experiment.

Table 33
Descriptive Statistics for Transcript Density (a) and Lag Coherence (b) as a Function of Co-located and Distributed Collaboration for Experiments 3 and 4

(a)	Co-located Density		Distributed Density	
	Exp 3	Exp 4	Exp 3	Exp 4
N	34	29	51	31
Mean	0.057	0.058	0.058	0.056
SD	0.006	0.004	0.004	0.007

(b)	Co-located Lag Coher		Distributed Lag Coher	
	Exp 3	Exp 4	Exp 3	Exp 4
N	34	44	51	41
Mean	-0.167	-0.158	-0.187	-0.133
SD	0.046	0.046	0.075	0.041

One-way ANOVAS were run with the two-level team distribution as the factor variable and communication metric as a dependent variable. These results are given in Table 34. These results indicated non-significant effects for Experiment 3, but marginal (density) and significant (lag coherence) effects for Experiment 4. In this experiment, co-located teams exhibited greater transcript density and larger lag coherence slopes (although not statistically detectable, it is interesting to note that the opposite pattern occurred in Experiment 3).

Table 34
F-statistics for Team Distribution ANOVAs for Transcript Density and Lag Coherence for Experiments 3 and 4

Factor	Experiment 3 Density	Lag Coher
Distribution	$F(1,83) = .073$	$F(1,83) = 1.937$
	Experiment 4 Density	Lag Coher
Distribution	$F(1,58) = 2.617$*	$F(1,83) = 6.625$**

Note. * $p < .15$; ** $p < .05$

A 2 (distribution condition) X 2 (level of workload) ANCOVA with team performance as the dependent variable and communication measure as the covariate was run in order to investigate H4.1. (Workload was used to control for a significant source of variance. Refer to task 4.) For Experiment 3, the interaction between density and distribution condition and lag coherence and distribution condition were both non-significant (F (1, 79) = .025 and F (1,79) = .001, respectively). Therefore for density and lag coherence measures in Experiment 3 there was no support for H4.1 that the communication – performance relationship would be greater for the distributed condition.

6.6.2.3 Co-located vs. distributed flow effects. *CHUMS findings.* Distributed teams had more CHUMS models than co-located teams for Missions 2, 4, and 5 (respectively $t(89) = 2.08, p = .04$; $t(89) = 3.00, p = .003$; $t(89) = 2.95, p = .004$). The number of models increased from Mission 5 to 6 ($t(89) = 2.03, p = .045$), particularly for co-located teams. This means three things. First, it means that distributed teams had less stable communication patterns than did co-located teams. Second, it means that, at the communication glitch (breakdown in communication between AVO and DEMPC in one direction in Mission 6), all teams exhibit an increase in communication patterns, and hence a decrease in stability. This is presumably due to adaptation to the glitch. Third, since co-located teams had a larger increase in CHUMS models than did distributed teams, and since distributed teams had more models to begin with, these findings show that, at the glitch, all teams behave more like distributed teams. This claim is supported by other findings as well. Co-location's impact on model counts drops and reverses between Missions 5 and the glitch mission (6) ($t(89) = -2.61, p = .011$), and the reversal becomes more extreme from 6 to 7 ($t(89) = -2.62, p = .010$). Overall, distributed teams tried more communication strategies than did co-located teams, though co-located teams also used more strategies at the glitch. Geographic distribution and a glitch seem to be associated with more patterns which may be attributed to team adaptation.

Dominance findings. In terms of Dominance, AVO is reactive in Co-located teams, but dominant in distributed teams ($F(1, 16.78) = 16.41, p = .001$). DEMPC is moderately

dominant for Co-located, but reactive for Distributed, $F(1, 16.51) = 12.84, p = .002$. During the Glitch Mission (6), Co-located teams become AVO-dominant ($t(89) = -2.08, p = .040$), and PLO-reactive ($t(89) = -1.88, p = .063$). Averaging across both conditions, DEMPC is dominant in the Glitch Mission (6) ($t(89) = 2.05, p = .043$), but reactive in all other missions. As with CHUMS findings, we see here that all teams behave more like distributed teams, during the glitch.

ProNet findings. ProNet patterns in Experiment 3 revealed a fairly straightforward picture. In Co-located teams (but not in Distributed teams) DEMPCs interrupt AVOs, and AVOs repeat themselves. Also, DEMPCs pass information to PLOs, and PLOs pass information to AVOs. In terms of two types of interruption, AiD and PiA, the impact of co-location decreases over time, after teams have reached performance asymptote, supporting H4.3.

For Experiment 4, ProNet findings revealed that co-located teams had more open communication channels than distributed teams. As with Experiment 3, most communication patterns were more common for co-located than for distributed teams, presumably because co-located teams spent more time talking than distributed teams. However, distributed teams had more APcycles than Co-Located teams. This suggests that AVO and PLO talked to one another more in distributed teams, presumably because DEMPC was more remote.

Turning to glitch effects for ProNet, between Missions 5 and 6, DAcycles decrease ($Wald(1) = 3.15, p = .076$), DPcycles increase ($t(89) = 1.82, p = .073$), and PAcycles increase ($t(89) = 2.11, p = .038$). Also, PDcycles decrease ($Wald(1) = 3.30, p = .069$). These behaviors simply validate that teams did, in fact, tend to adjust their communication patterns during the glitch mission. Since DEMPC was unable talk to AVO, teams developed additional patterns, in which DEMPC-to-AVO communication decreases, but DEMPC-to-PLO communication increases, as does PLO-to-AVO. Given that we cut the communication channel, these findings are not particularly surprising in and of themselves. It simply means that teams did in fact adapt. More importantly, it means that we were able to detect and diagnose this adaptation, with our communication measures.

CRP findings. Distributed teams tended to have higher PLO CRP values, indicating that distributed PLO's tended to speak in more norm-appropriate ratios than did co-located PLO's. This may be due to AVO and DEMPC talking too much in co-located teams. However, this Distributed communication advantage decreases with task experience.

Hypothesis 4.1 was that the performance-communication relationships identified under Task 3 would be stronger in the distributed condition, than the co-located condition. This hypothesis was not supported. Findings for the distributed condition did not match the general performance-FAUCET relationship. For example, the CHUMS finding for performance was that high performing teams have fewer models. However, the co-location finding was that distributed teams had more CHUMS models than did co-located teams, and yet they performed about as well.

In summary, though distributed teams spent less total time speaking, they tended to use more words per second (see the next section). This means that they speak more quickly, and/or use shorter words. CHUMS measures showed that distributed teams have more models, and hence have more distinct communication patterns, than do co-located teams. This reveals that their communication pattern is less stable than for co-located teams. Taken together, these findings suggest that, compared to co-located teams, distributed teams speak less, but get to the point more quickly. They do not develop strong formalisms for communication style. Hence, though they use more words, it is likely that the words they use are more function-oriented. This can be interpreted as a more terse communication style.

6.6.2.4 Differences in amount of talking and keywords. Additional methods were developed and tested to use basic textual features to predict whether teams were co-located or distributed based on their discourse. Aspects of this novel methodology might be useful in other team situations in terms of characterizing language differences among teams.

For example, Table 35 shows the results of correlating the total number of words in a team-at-mission transcript with the length of the transcript in time. There was a significant difference in the correlations(r difference $p<.05$). This suggests that the discourse for distributed teams may be more uniformly distributed during the mission. Other results under this task confirm this hypothesis. Both of these correlations are significantly different from 0 ($p<.01$).

Table 35
Correlation of Number of Words with Transcript Time

Team Type	Correlation
All	0.362
Co-located	0.285
Distributed	0.530

The hypothesis that distributed teams will speak more (H4.2), since all communication must be spoken is supported by our findings in Table 36 showing that distributed team average significantly more words per second than co-located teams (t (168)=5.06, $p<.01$).

Table 36
Average Words Per Second

Team Type	Average Words Per Second
Co-located	0.869
Distributed	1.069

In what we believe is a useful approach to detecting team differences based on dialogue, we investigated the ratios of the frequencies of common words and phrases for co-located and distributed teams. We computed these for words (unigrams), two word sequences (bigrams) and three word sequences (trigrams). This procedure is similar to that of the KWI analysis.

Our algorithm computes the most frequent words (bigrams, trigram respectively) for each co-located (respectively distributed) team-at-mission transcript. The 200 most frequent words from all co-located (respectively distributed) teams is compiled and frequency of occurrences of the word over all co-located (respectively distributed) transcripts is computed. We then compute the ratio of the frequencies of the co-located team's words (respectively bigram, trigrams) to the frequencies of those for the distributed teams and vice versa. The results, sorted by ratio, are shown in the three tables below (Tables 37, 38, and 39). While we believe this is a useful approach, we have not been able to reach any definite conclusions from our analyses of this data. Our working hypothesis that the discourse for distributed teams would be more focused and "on-task" the results below neither confirm nor rule out the hypothesis.

Table 37
Ratios of Most Frequent Unigrams

Word	Di-Freq	Co-Freq	Ratio (Di/Co)
lej	116	32	6.378993083
required	66	20	5.807083358
stats	39	12	5.719097247
holding	61	19	5.649634527
regulations	32	10	5.631111135
steady	131	43	5.361014235
ork	78	26	5.279166689

Word	Co-Freq	Di-Freq	Ratio (Co/Di)
gonna	549	2	155.9905
de	131	2	37.22178
advise	119	2	33.81215
dem	58	1	32.95975
wlf	83	2	23.58327
gotta	36	1	20.45777
inform	95	3	17.99526

mountain	38	13	5.143803441
consensus	21	8	4.619270853
cued	73	28	4.587847242
who	17	7	4.273611129
immediately	21	9	4.106018536
ste	51	22	4.079356078
actual	18	8	3.959375017
red	75	34	3.881740213
her	89	41	3.81988484
she's	8	4	3.519444459
mst	16	8	3.519444459
weird	10	5	3.519444459
target's	26	13	3.519444459
picture's	53	28	3.330902792
sec	15	8	3.299479181
mkl	60	32	3.299479181
hundred	28	15	3.284814829
jump	13	7	3.268055569
side	13	7	3.268055569
bunch	9	5	3.167500013
confused	9	5	3.167500013
kidding	9	5	3.167500013
man	43	24	3.152835662
non	16	9	3.128395075
exiting	7	4	3.079513902
here's	7	4	3.079513902
safe	14	8	3.079513902

mar	63	2	17.90055
x	27	1	15.34333
elapsed	24	1	13.63852
noted	24	1	13.63852
y	24	1	13.63852
photographed	60	3	11.36543
rain	19	1	10.79716
tke	74	4	10.51302
check	526	29	10.30727
present	18	1	10.22889
intercom	16	1	9.092344
op	16	1	9.092344
overshot	15	1	8.524073
recommendations	15	1	8.524073
slt	14	1	7.955801
pilot	54	4	7.671665
calling	39	3	7.38753
photographing	13	1	7.38753
stopped	13	1	7.38753
primary	26	2	7.38753
reading	25	2	7.103394
ssrt	12	1	6.819258
cancel	11	1	6.250987
flaps	11	1	6.250987
whether	11	1	6.250987
wp	11	1	6.250987
excellent	55	5	6.250987

without	14	8	3.079513902
zero	14	8	3.079513902
sen	306	175	3.077000013
[garbled]	310	178	3.064684782
one's	12	7	3.01666668
forgot	15	9	2.932870383
joh	15	0	0
shit	15	0	0
buddy	16	0	0
cah	16	0	0
cert	16	0	0
fart	17	0	0
haul	17	0	0
intended	17	0	0
sitting	17	0	0
reg	18	0	0
regs	18	0	0
narrow	21	0	0
romeo	21	0	0
that'd	22	0	0
dash	23	0	0
include	23	0	0
mark	23	0	0
bump	27	0	0
gre	27	0	0
photographable	34	0	0
umm	34	0	0
regulation	39	0	0
thousand	45	0	0
re-enter	48	0	0
proposed	49	0	0
storm	49	0	0
letting	51	0	0
er	52	0	0
smi	55	0	0
dmpc	82	0	0

communicate	21	2	5.966851
communication	21	2	5.966851
maintained	21	2	5.966851
affective	315	30	5.966851
bypass	10	1	5.682715
intel	10	1	5.682715
adjustment	14	0	0
dave	14	0	0
experimentor	14	0	0
queued	15	0	0
terrain	15	0	0
effect	17	0	0
duke	18	0	0
ban	19	0	0
letters	19	0	0
fstr	22	0	0
expected	23	0	0
cooter	24	0	0
bay	30	0	0
fste	33	0	0
coordinate	36	0	0
q	36	0	0
standing	36	0	0
barea	44	0	0
kob	45	0	0
successfully	46	0	0
luv	48	0	0
van	53	0	0
successful	61	0	0
sur	83	0	0
agt	88	0	0
san	91	0	0
nha	106	0	0
crew	115	0	0
harea	415	0	0
farea	478	0	0

Table 38
Ratios of Most Frequent Bigrams

Word	Di-Freq	CoFreq	Ratio (Di/Co)
that shot	27	1	37.54096591
accepted proceed	20	1	27.8081229
proceed roger	19	1	26.41771675
special requirements	19	1	26.41771675
a mountain	15	1	20.85609217
straight and	30	2	20.85609217
photo let's	29	2	20.1608891
altitude be	14	1	19.46568603
for red	14	1	19.46568603
okay [null]	14	1	19.46568603
your intended	14	1	19.46568603
for ork	13	1	18.07527988
get sen	13	1	18.07527988
pictures uh	13	1	18.07527988
regulation is	13	1	18.07527988
we're sitting	13	1	18.07527988
will your	25	2	17.38007681
only an	12	1	16.68487374
picture's going	12	1	16.68487374
exit at	23	2	15.98967067
2 i	11	1	15.29446759
above uh	11	1	15.29446759
cause i'm	11	1	15.29446759
regulations on	11	1	15.29446759
guys after	22	2	15.294467

Word	Co-Freq	Di-Freq	Ratio (Co/Di)
gonna be	101	1	72.40535531
please advise	96	1	68.82093178
dempc to	141	2	50.54037178
clear picture	54	1	38.71177413
check on	252	5	36.13098919
okay thank	45	1	32.25981177
clear to	44	1	31.54292707
the cue	42	1	30.10915765
will proceed	42	1	30.10915765
plan on	41	1	29.39227295
have good	40	1	28.67538824
acceptable picture	38	1	27.24161883
speed min	38	1	27.24161883
entry site	36	1	25.80784942
of 5.0	35	1	25.09096471
and switch	32	1	22.94031059
over all	30	1	21.50654118
to wlf	29	1	20.78965648
right roger	86	3	20.55069491
i'm back	28	1	20.07277177
picture looks	28	1	20.07277177
to sel	28	1	20.07277177
of at	27	1	19.35588706
ok thank	27	1	19.35588706
roger we	53	2	18.997444

			59
restrictions only	21	2	14.59926452
altitude regulation	10	1	13.90406145
changed course	10	1	13.90406145
did say	10	1	13.90406145
give them	10	1	13.90406145
heading after	10	1	13.90406145
in it	10	1	13.90406145
mste um	10	1	13.90406145
requirements just	10	1	13.90406145
speed holding	10	1	13.90406145
uh sstr	10	1	13.90406145
uh 3000	20	2	13.90406145
holding steady	39	4	13.55645991
a storm	19	2	13.20885838
photo target	19	2	13.20885838
charlie romeo	15	0	0
dmpc do	15	0	0
speed 100-200	15	0	0
we re-enter	15	0	0
3,000 okay	16	0	0
include speed	16	0	0
regulation of	16	0	0
now changed	17	0	0
shot off	17	0	0
the er	17	0	0
your proposed	17	0	0
and narrow	21	0	0
to re-enter	21	0	0
3,000 feet	22	0	0
photographable target	24	0	0
camera target	26	0	0
dmpc this	26	0	0
proposed air	34	0	0
re-enter at	35	0	0

			71
move to	26	1	18.63900236
200 knots	156	6	18.63900236
my radius	51	2	18.28056
be photographed	25	1	17.92211765
the coordinates	25	1	17.92211765
a clear	97	4	17.38445412
a total	24	1	17.20523295
dempc be	24	1	17.20523295
for photo	24	1	17.20523295
total of	24	1	17.20523295
knots roger	23	1	16.48834824
me what's	23	1	16.48834824
remaining and	23	1	16.48834824
target sstr	23	1	16.48834824
dempc roger	45	2	16.12990589
advised that	39	0	0
successful picture	41	0	0
after farea	43	0	0
are clear	43	0	0
10-4 avo	45	0	0
am gonna	46	0	0
crew plo	46	0	0
miles over	47	0	0
de please	53	0	0
for harea	57	0	0
change over	59	0	0
to way	59	0	0
uh de	65	0	0
cue in	71	0	0
gonna go	84	0	0
to farea	92	0	0
to harea	97	0	0
for farea	100	0	0
we're gonna	115	0	0

above 3,000	40	0	0	i'm gonna	151	0	0

Table 39
Ratios of Most Frequent Trigrams

Top 50 3-gram for Collocated			
ngram	Co	Dist	Ratio
is 5 0	231	9	25.66667
to avo i	90	4	22.5
copy that plo	65	3	21.66667
okay i am	62	3	20.66667
its effective radius	80	4	20
the picture plo	55	3	18.33333
of 5 0	128	7	18.28571
effective radius will	48	3	16
5 0 roger	42	3	14
miles copy that	38	3	12.66667
next picture will	38	3	12.66667
radius will be	63	5	12.6
5 0 and	37	3	12.33333
and max of	36	3	12
next target area	48	4	12
on that okay	48	4	12
plo we have	84	7	12
have a minimum	47	4	11.75
yes that s	35	3	11.66667
ahead and cue	45	4	11.25
is dempc i	90	8	11.25
affirmative thank you	33	3	11
five mile radius	54	5	10.8
5 it is	32	3	10.66667
also a target	32	3	10.66667
affective radius is	128	12	10.66667
you tell me	128	12	10.66667
to 400 knots	42	4	10.5
of 2 50	31	3	10.33333
within five miles	31	3	10.33333
that go ahead	41	4	10.25
next target roger	30	3	10
next point after	108	11	9.818182
is dempc you	49	5	9.8
are my restrictions	29	3	9.666667
can you tell	116	12	9.666667
plo how many	48	5	9.6
will be a	48	5	9.6
five miles the	28	3	9.333333
have we taken	28	3	9.333333

Top 50 3-gram for Distributed			
ngram	Dist	Co	Ratio
just letting you	43	4	10.75
this is uh	27	3	9
uh no restrictions	26	3	8.666667
going to exit	34	4	8.5
need you at	41	5	8.2
and we do	23	3	7.666667
letting you know	43	6	7.166667
good for the	28	4	7
onto next target	28	4	7
must be above	20	3	6.666667
is a negative	19	3	6.333333
next target uh	19	3	6.333333
uh dempc this	24	4	6
dempc no restrictions	23	4	5.75
what will be	57	10	5.7
uh effective radius	34	6	5.666667
the maximum is	28	5	5.6
no speed or	22	4	5.5
uh speed rules	22	4	5.5
guys we got	16	3	5.333333
point it s	16	3	5.333333
that we got	37	7	5.285714
uh what do	21	4	5.25
1000 to 3000	15	3	5
5 0 speed	15	3	5
a exit point	15	3	5
picture s taken	15	3	5
we ll hit	15	3	5
re out of	25	5	5
f area avo	24	5	4.8
i ll need	19	4	4.75
0 and we	14	3	4.666667
ready to move	14	3	4.666667
site is going	14	3	4.666667
uh dempc what	28	6	4.666667
what would be	18	4	4.5
3000 effective radius	13	3	4.333333
uh f area	13	3	4.333333
you must be	30	7	4.285714
avo are there	51	12	4.25

i m sending	28	3	9.333333
the picture let	28	3	9.333333
avo dempc this	37	4	9.25
avo i got	37	4	9.25
3 75 miles	27	3	9
a photograph of	27	3	9
five miles for	27	3	9
roger avo this	27	3	9
sen 2 roger	27	3	9
speed 300 maximum	27	3	9

100 and 200	21	5	4.2
5 that s	21	5	4.2
need you below	21	5	4.2
this is the	126	31	4.064516
give me one	12	3	4
have any flight	12	3	4
is taken and	12	3	4
must be under	12	3	4
you are correct	12	3	4
you need is	12	3	4

6.6.3 Conclusions

- Co-located teams exhibited greater transcript density (content efficiency) and larger lag coherence slopes (stability) though these results did not replicate across experiments
- Distributed teams are more variable (i.e., more CHUMS and ProNet patterns) than co-located
- At the communication glitch, all teams tend to behave as if distributed indicating that increased patterns may be a form of team adaptation
- ProNet methods detect team's adaptation to the communication channel glitch. As expected, teams reroute communication to avoid the cut channel.
- Co-located teams had more open communication channels than distributed
- Distributed PLO's tended to speak only as much as they needed to, while co-located PLO's tended to speak more frequently than needed.
- The hypothesis that distributed teams will speak more (i.e., more words per minute), since all communication must be spoken is supported
- Discourse for distributed teams may be more uniformly distributed during the mission.
- Compared to co-located teams, distributed teams take spend less time speaking though use more words when they do, have more distinct, but less stable communication patterns. Distributed teams are terse, but with no strong formalisms.

We again note that consistency is key as performance develops, but less so for distributed teams, as they seem to have a proclivity for more and varied patterns. While tentative, these results empirically support what has often been inferred by studies in the team cognition literature, that expectancies and thus consistencies develop over time and are the hallmark of highly skilled teams. Interestingly, distributed teams, as opposed to co-located teams, are less consistent in terms of communication patterns, suggesting that the distributed environment (as well as the glitch) increases adaptation behavior on the part of the team. Our flow analysis methods have also demonstrated their utility for uncovering communication glitches (i.e., a breakdown in communication equipment). These can be indexed either by a sudden change in pattern, or in who is atypically leading

the flow of conversation. Important for this grant's work, this indexes differentially based on whether the team is co-located or distributed.

6.7 Task 5: Examine Impact of Workload on Communication and Performance

6.7.1 Method

In both experiments, workload was manipulated within teams. In Experiment 3, the last 3 missions of seven were high workload and in Experiment 4, the last mission of five was high workload. It is not clear how this workload manipulation should affect team communication. How do communication patterns change with increasing workload and with associated performance decrements? Is the impact of the communication breakdown moderated by the environment (i.e. distributed vs. co-located)? These questions will be addressed by examining the results of our communication metrics by workload condition

Initial analyses include examining communication across workload conditions. If differences in communication metrics are detected the nature of the differences are more fully explored.

We hypothesized based on our previous studies and those in the literature that teams would tend to speak less under high workload (H5.1) and would use more action-directed language (H5.2). This was expected to hold more for effective (i.e., better performing) teams than ineffective teams (H5.3). Moreover, effective teams were expected to remain taskwork-focused more than poor performing teams, because their teamwork patterns had been established by the time workload increased. In addition, we anticipated that teams would employ alternate communication paths when confronted with the glitch. Effective teams were expected to have a more efficient teamwork routine developed, and so should have been faster to switch into and out of the alternative communication paths employed during the glitch (H5.4). Unfortunately these analyses presupposed the capability to determine the precise timing of the glitch which was event-based, not temporally-based. Therefore, the information on timing of the glitch was not available making it impossible to address this hypothesis. Finally, in terms of content coherence, we expected that during early missions and under low workload, high-performing teams would also have higher coherence ratings than low-performing teams because they are communicating the task information clearly. However, during the later high workload mission, higher-performing teams should have lower coherence scores by virtue of the fact that they shared expectations of the task and so can afford to reduce explicit communication (H5.5).

Table 40
Hypotheses Associated with Task 5

Task/Hypothesis Number	Hypothesis	Supported?
5.1	Teams will communicate less under high workload than low.	Yes
5.2	LSA-based coding will indicate more action-oriented communications under high workload compared to low.	?
5.3	The patterns in 5.1 and 5.2 are expected to hold more for high-performing teams than low.	Yes
5.4	High-performing teams should be faster to switch into and out of the alternative communication paths employed during communication breakdowns than low-performing teams.	?
5.5	Under low workload, high-performing teams should have high coherence scores, but under the later high workload missions, higher-performing teams should have lower coherence scores.	No

* There were no data available to test the H5.2 and H5.4

6.7.2 Results

6.7.2.1 Predicting workload level using whole transcripts. The goal of using LSA to predict workload involved modeling the transcripts as a whole for both high and low workload teams. By building models of the language used in high and low workload teams, one can then predict the workload level of any new team. Using LSA and whole transcripts we were able to accurately predict whether a team was under low or high workload conditions.

Using a similar k-nearest neighbor algorithm (as above) on whole transcripts to predict workload we found strong correlations between the actual and predicted workloads. The algorithm first assigns a score of 1 for high workload and 0 for low workload missions. Then it takes the average, weighted by distance in the semantic space, of the 10 closest team-at-missions, excluding all missions from the current team and from other experiments. "Team-at-missions" whose weighted average is greater than the cutoff of 0.25 are labeled "high workload," others are labeled "low workload." We computed the Kappa statistic to assess the agreement between the actual and predicted workload.

Experiment 3 Kappa = 0.91
Experiment 4 Kappa = 0.84

This result shows that the approach can accurately classify a mission as to whether the team was under high or low workload. (Recall this is Cohen's Kappa, a chance corrected measure of agreement between the predicted and actual workload labels.)

As an additional check, we redid the workload analyses taking out waypoint names so that the LSA analyses would not make workload predictions based on the specific waypoints mentioned, but instead on the actual content expressed during the missions. (Low and High Workload conditions differed by the number of target waypoints and thus there were unique identifiers in the form of target waypoint names associated with High, but not Low Workload.)

In this analysis, we predicted high (1) vs. low (0) workload using the LSA-based 10 closest measure omitting all waypoint names from the transcripts. A discriminant function was derived from the LSA measure in order to determine how well the measure could distinguish transcripts based on workload. For the Experiment 3 transcripts, the function classified 91.8% of the transcripts correctly (90.6% with cross-validation) and for the Experiment 4 transcripts, it classified 96.5% of the transcripts correctly (96.5% with cross-validation). The results suggest that LSA is able to easily distinguish workload based on team language characteristics, even when the specific mission information is not present. However, it is still possible that LSA is not picking up on content differences in Low vs. High Workload discourse, but rather the amount of speaking (in the vector length metric) which is lower for Low Workload than High.

6.7.2.2 Varying LSA parameters to predict team workload. The graph below shows the results of varying parameters in the correlation of predicted and actual team workload. With most parameters there was a strong correlation, indicating the predictive value of our model. The parameters varied are the radius and the corpus of team-at-mission transcripts used to make the prediction. The radius is the number of "closest" teams used to make the prediction. For example, when the radii of the ten closest teams (based on their whole transcript discourse as assessed by LSA) are chosen, the workload level of each of the ten teams is looked up and a weighted average is computed to assign the predicted workload level. Ten is a standard choice for this type of task and appears to work best overall in this case.

The other parameter we varied was the composition of the corpus from which the R (radius) closest teams were chosen. In Figure 10 **Exp_AFx** indicates all team-at-mission transcripts from the same experiment were excluded from the corpus. **Mission_AFx** indicates only the current team-at-mission transcript was excluded from the corpus. **Team_AFx** indicates all team-at-mission transcripts for the same team were excluded from the corpus. **TExOExp_AFx** indicates that all team-at-mission transcripts for the same team and all team-at-mission transcripts from other experiments were excluded from the corpus. TExOExp_AFx is the most commonly used value for this parameter followed by Team_AFx, because it is generally best to exclude all other team-at-mission transcripts from the team for which we are making a prediction to avoid overfitting our model to the data. Our results below show good predictive performance for both of these parameter values on the workload prediction task.

These results show that the measures, including the most conservative, still provide highly accurate predictions of whether teams were in high or low workload conditions. This suggests that this approach can be incorporated into systems which could predict if a team's workload has suddenly changed, or is getting to a level that may cause deterioration in performance.

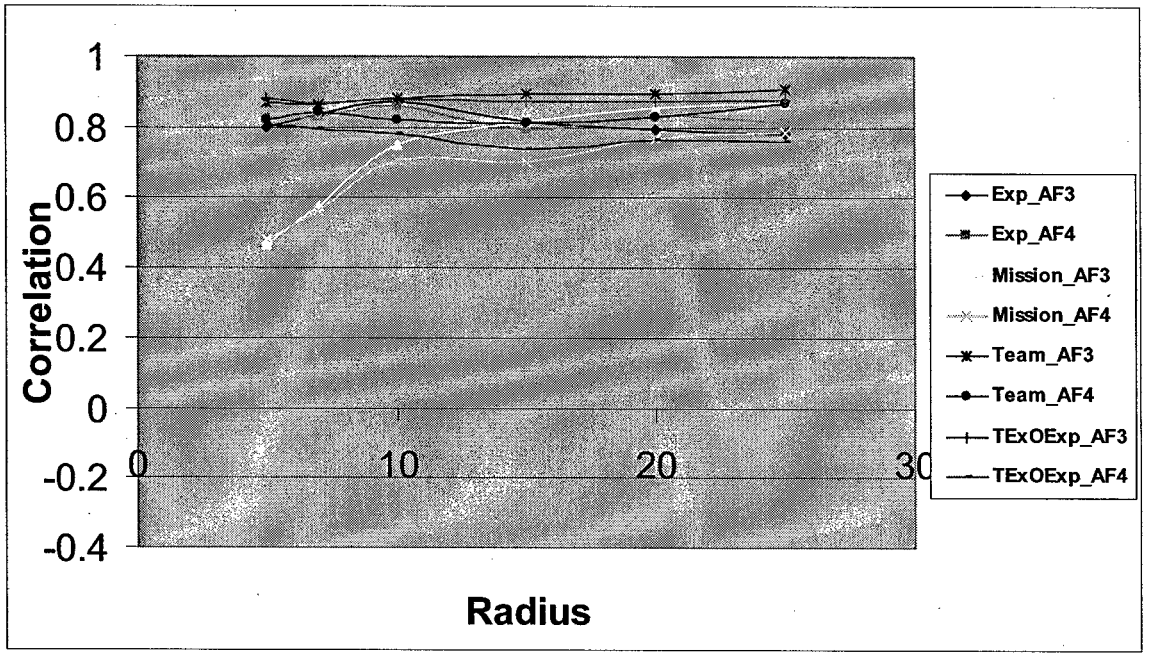


Figure 10. Work-Load Predictions Correlations Using Experiments 1, 3, and 4.

6.7.2.3 Workload Effects on transcript density and lag coherence. The LSA-based density and lag coherence measures were analyzed for effects of high workload in Experiments 3 and 4. Table 41 lists the number of observations, means, and standard deviations of each score for each experiment.

Table 41
Descriptive Statistics for Transcript Density (a) and Lag Coherence (b) as a Function of Workload for Experiments 3 and 4

(a)	Low WL Density		High WL Density	
	Exp 3	Exp 4	Exp 3	Exp 4
N	45	46	40	14
Mean	0.057	0.056	0.058	0.059
SD	0.005	0.005	0.005	0.005

(b)	Low WL Lag Coher		High WL Lag Coher	
	Exp 3	Exp 4	Exp 3	Exp 4
N	45	68	40	17
Mean	-0.169	-0.148	-0.19	-0.134
SD	0.063	0.046	0.068	0.041

Four one-way ANOVAS were used with the two-levels of workload as the factor variable and each of the communication metrics as the dependent variable for each experiment. The results of these analyses are presented in Table 42. The only significant workload effect was on transcript density in Experiment 4. Specifically, transcripts exhibited greater density under high workload than low workload (a similar but non-significant pattern was observed for Experiment 3). Although H5.5 is not supported, the density finding partially supports H5.1. Teams pack more content-laden terminology into the same amount of discourse when under pressure from high workload.

Table 42
F-statistics for Workload ANOVAs for Transcript Density and Lag Coherence for Experiments 3 and 4

IV	Experiment 3	
	Density	Lag Coher
Workload	$F(1,83) = 1.313$	$F(1,83) = 2.143$
	Experiment 4	
	Density	Lag Coher
Workload	$F(1,58) = 3.052$*	$F(1,83) = .560$

Note. * $p < .10$

6.7.2.4 Effects of workload on communication flow. Using a variety of measures flow measures with amount of talking as a covariate, we found that teams talk less under high workload, supporting H5.1.

CHUMS findings. There were fewer CHUMS models per minute for models that include the null parameter, as well as for those without.

For CHUMS included null parameter models per minute:
$F(1, 15) = 12.49, p = 0.003, \eta^2 = 0.454$
Mission 4: $M = 0.275$, $SE = 0.008$
Mission 5: $M = 0.233$, $SE = 0.009$

For CHUMS excluded null parameter models per minute:
$F(1, 15) = 13.43, p = 0.002, \eta^2 = 0.472$
Mission 4: $M = 0.205$, $SE = 0.007$
Mission 5: $M = 0.169$, $SE = 0.007$

ProNet findings. This decrease in communication at high workload was also found with two of the ProNet measures. High workload was marked by a decrease in chain length sums. Also, the high workload mission yielded was less likely to yield a detectable A-P cycle than was the previous mission.

For ProNet chain length sums:
$F(1, 16) = 3.44, p = 0.082, \eta^2 = 0.177$
Mission 4: $M = 54.56$, $SE = 5.62$
Mission 5: $M = 39.38$, $SE = 5.95$

For ProNet A-P Cycles:
$F(1, 16) = 3.43, p = 0.083, \eta^2 = 0.176$
Mission 4: $M = 0.722$, $SE = 0.065$
Mission 5: $M = 0.546$, $SE = 0.069$

In fact, for the effect of ProNet chain length sums, this adaptation also leads to better performance ($F(1, 14) = 4.14, p = 0.061, \beta = -0.546$). This indicates that good teams adapt to high workload by reducing their utterance chain lengths, supporting H5.3.

Dominance findings. Findings were also uncovered for dominance. At the high workload mission, AVO becomes more dominant, and DEMPC becomes more reactive. This pattern is closer to the normal pattern for distributed teams. Hence, increased workload, like the communication glitch, causes teams to behave more like distributed teams.

For AVO:
$F(1, 16) = 6.05, p = 0.026, \eta^2 = .275$
Mission 4: $M = -0.417$, $SE = 0.124$
Mission 5: $M = 0.028$, $SE = 0.131$

For DEMPC:
$F(1, 16) = 6.36, p = 0.023, \eta^2 = .285$
Mission 4: $M = 0.272, SE = 0.150$
Mission 5: $M = -0.279, SE = 0.159$

CRP findings. CRP-based measures did not reveal any meaningful workload effects.

6.7.3 Conclusions

- An LSA-based approach can accurately classify a mission as to whether the team was under high or low workload.
- LSA-based workload classification findings are robust over variations in the radius and the corpus of team-at-mission transcripts used to make the prediction
- Transcripts exhibited greater density under high workload than low workload
- Teams speak less under high workload, supporting H5.1
- Under high workload, AVO becomes more dominant, and DEMPC becomes less so, resembling the patterns seen in distributed teams
- Under high workload, all teams behave more like distributed teams
- Good teams adapt to high workload by reducing their utterance chain lengths, supporting H5.3.

6.8 Task 6: Investigate Link Between Communication and Shared Mental Models

6.8.1 Method

Teamwork and taskwork knowledge were measured in both experiments (twice in Experiment 3, once in Experiment 4. Under Task 6, we investigate the relationship between team communication and SMMs (shared mental models). We computed accuracy of SMMs in two ways, holistic or collective knowledge accuracy. For taskwork knowledge this involved averaging individual knowledge scores based on comparing individual Pathfinder networks to individual referent networks (i.e., one for each team member and one for the team as a whole), where accuracy was proportion of shared links. Note that overall knowledge accuracy is scored at the individual level against a team-level referent. For holistic taskwork knowledge, a consensus Pathfinder network (i.e., one generated by consensus on each pairwise rating) was compared to the team-level referent in terms of proportion shared links. Individual teamwork accuracy involved comparing individual responses on a teamwork knowledge questionnaire to an expert-response metric. These teamwork scores were combined through averaging (i.e., collective score). Team members were also asked to come to consensus on responses to the teamwork questionnaire. This consensual response was also compared to the expert referent. This results in a total of four SMM Accuracy Scores for each team in the two experiments: collective taskwork accuracy, holistic taskwork accuracy, collective

teamwork accuracy, and holistic teamwork accuracy. For the results presented here, each of these measures is based on team member knowledge at the end of the experiments.

We also looked at taskwork and teamwork SMMs in terms of intrateam similarity. Similarity scores were derived for each pair of team members on a team for teamwork or taskwork knowledge. These similarity scores were averaged across pairs within a team. We refer to these averages as SMM Similarity Scores.

How do differences along this knowledge sharing dimension correspond to communication patterns? In this section we look for relationships (i.e., correlations) between our communication metrics and the knowledge scores. We also examine the impact of the environment (distributed vs. co-located) on knowledge sharing and associated communication. It is possible that certain specific communication patterns are critical for the development of shared mental models in distributed environments, but less important for co-located teams. For instance, distributed teams may require more feedback on team member actions, roles, or plans. Consistency of relationships across experiments will be examined.

We made several predictions. Specific predictions include how sequential flow will differ based on variations in knowledge scores. Specifically, teams with more taskwork knowledge should have longer mean chain lengths because they have an ordered and rehearsed communication pattern (H6.1). Similarly teams with high levels of teamwork knowledge should exhibit more stability in terms of communication flow because they will have better established how much each team member should speak during each phase of the task (H6.2). We also predict based on previous studies that increased knowledge sharing (in this task demonstrated by increases in interpositional knowledge and intrateam similarity) should correspond to decreases in communication frequency (H6.3). Finally we predict that individual team members with higher overall taskwork and teamwork knowledge scores should also have higher dominance scores, relative to those with lower knowledge.

Table 43
Hypotheses Associated with Task 6.

Task/Hypothesis Number	Hypothesis	Supported?
6.1	Teams with more taskwork knowledge should have longer mean chain lengths representing communication flow.	No
6.2	Teams with high levels of teamwork knowledge should exhibit more stability in terms of communication flow.	No
6.3	Increased interpositional knowledge and intrateam similarity should correspond to decreases in communication frequency.	Partial
6.4	Individual team members who understand the task the best will have higher dominance scores.	Yes

6.8.2 Results

6.8.2.1 Transcript density, lag coherence, and shared mental models. An exploratory approach was undertaken in order to identify any relationships between SMM Accuracy Scores and team communication content measures. Specifically, all mission Transcript Density and Lag Coherence scores were evaluated for linear correlation with the SMM measures. The results (over both experiments) indicated that collective teamwork knowledge was most consistently correlated with the content metrics (i.e., other SMM Accuracy measures correlated sporadically with various missions, but were not nearly as consistent as collective teamwork knowledge).

In order to further investigate the relationship between team communication content and SMMs, we used Experiment 3 (Knowledge Session 2) and Experiment 4 taskwork SMM Similarity Score as a covariate with team distribution condition (in the following all workload effects were *ns*) to predict transcript density and lag coherence. In other words we looked at the relationship between team members' shared mental models of the task (degree of sharing) and communication content. In terms of the taskwork SMM Similarity Scores, only the transcript density scores seemed to be related (all lag coherence effects ns). For Experiment 3, there was a condition by covariate interaction ($F(1, 76) = 29.97, p < .001$), suggesting that the relationship between taskwork SMM Similarity Scores and transcript density were different for co-located and distributed teams. Namely, this relationship was negative for co-located teams ($t(29) = -5.77, p < .001$), but nonexistent (*ns*) for distributed teams. The same analysis on Experiment 4 data revealed a significant relationship between taskwork SMM Similarity Score and transcript density ($F(1, 54) = 2.70, p = .10$) across team distribution conditions. *Post hoc* analysis revealed that this relationship was negative for each condition taken individually, but more so for co-located ($t(27) = -1.73, p < .10$) than for distributed ($t(29) = -1.52$, *ns*). Together, these results suggest that taskwork SMM Similarity Scores are more directly

related to certain aspects of team communication for co-located, compared to distributed teams. Further, this relationship tends to be negative with a higher degree of knowledge similarity related to lower degrees of transcript density, or the propensity to use less task-specific language partially supporting H6.3

The procedure previously applied to taskwork knowledge was also applied to teamwork knowledge to investigate the relationship between team communication content and teamwork SMMs. We used Experiment 3 (Knowledge Session 2) and Experiment 4 teamwork SMM Similarity Score as a covariate with team distribution condition (in the following all workload effects were *ns*) to predict transcript density and lag coherence. For Experiment 3, there was a significant distribution condition by Teamwork SMM Similarity Score for both lag coherence (F (1, 79) = 7.80, $p < .01$) and transcript density (F (1, 79) = 4.69, $p < .05$). A main effect of the covariate, teamwork SMM, was also found for lag coherence (t (79) = 5.77, $p < .001$). *Post hoc* analysis on the Experiment 3 lag coherence scores revealed that the interaction effect was due to a strong, positive relationship between Teamwork SMM Similarity and distributed team lag coherence (t (49) = 6.93, $p < .001$) and no relationship between Teamwork SMM Similarity Score and co-located team lag coherence (*ns*). This analysis suggests that the teamwork SMM similarity main effect on lag coherence was likely due to the distributed teams. *Post hoc* analysis on Experiment 3 transcript density interaction revealed a strong, negative relationship between Teamwork SMM Similarity Score and transcript density for distributed teams (t (32) = -4.43, $p < .001$), but no relationship for co-located teams (*ns*). Together these results indicate that the development of a teamwork SMM may be strongly and positively tied to the length of time co-located teams spend on a particular topic (lag coherence) but negatively tied to the use of task-specific language (transcript density) for distributed teams. The latter, but not the former result supports H6.3. These results were not replicated in the Experiment 4 data.

6.8.2.2 Shared mental models and communication flow. As was done for the analyses in Section 6.8.2.1, we examined four SMM Accuracy measures for each team in the two experiments: collective taskwork accuracy, holistic taskwork accuracy, collective teamwork accuracy, and holistic teamwork accuracy. For the results presented here, each of these measures is based on team member knowledge at the end of the experiments. In all cases of testing for a relationship between SMMs and FAUCET measures, we first included total amount of speech as a covariate. This is done to determine how much FAUCET measures contribute to knowledge, beyond the very basic measure of how much the team was talking.

For both Experiment 3 and Experiment 4, we found no FAUCET relationships between knowledge accuracy measures and CHUMS or ProNet. Therefore there is not support for H6.1 and H6.2. There were no CRP relationships for Experiment 4. However, for Experiment 3, holistic teamwork knowledge was positively correlated with PLO's CRP score, $t(12) = 2.561$, $p = .025$, *partial-r* = .595. This can be interpreted as follows: teams with greater teamwork knowledge also have PLO's that speak as normatively predicted. This pattern is similar to the findings for the performance measure, in which teams tend to perform better if PLO is an independent agent.

Dominance was an adequate predictor of all four knowledge accuracy measures in Experiment 3, and in three of the four for Experiment 4. In all cases, the predictor was the mean squared cross-correlation function between a pair of speakers, represented as ccf2xy (for speaker x to speaker y).

In regard to the Dominance metric results, let's first consider shared mental model accuracy for taskwork. For holistic taskwork, ccf2da was a negative predictor in Experiment 3 ($t(12) = -1.999, p = .069$, *partial-r* = -0.500). Ccf2dp was a negative predictor in Experiment 4 ($t(15) = -2.002, p = .064$, *partial-r* = -0.459). For collective taskwork at Experiment 3, ccf2pa was a negative predictor ($t(13) = -2.019, p = .065$, *partial-r* = -0.489). Finally, for collective taskwork at Experiment 4, ccf2ap was a positive predictor ($t(15) = 1.910, p = .076$, *partial-r* = .442). These are four disparate findings for taskwork. However, three of the correlations are negative, suggesting that every team member's taskwork is best developed when the team member is independent (in regard to communication flow), with one exception. Teams have better collective taskwork knowledge accuracy when PLO reacts to AVO's utterances. Generally, these results support H6.4 in that dominance is associated with high taskwork knowledge accuracy.

Turning to results of the Dominance metric for shared mental models of teamwork, we see a more consistent relationship for Experiment 3 between holistic and collective measures of SMM accuracy. For both holistic and collective teamwork, ccf2da and ccf2ad were required to form a predictive model, and in all cases, the relationship was positive. For holistic teamwork, ccf2ad was a non-predictive covariate, and ccf2da was positively related, $t(13) = 1.897, p = .080$, *partial-r* = .444. For collective teamwork, a positive relationship was found for both ccf2ad ($t(13) = 2.315, p = .038$, *partial-r* = .471), and for ccf2da ($t(13) = 2.133, p = .053$, *partial-r* = .431). Finally, for Experiment 4 collective teamwork, a positive relationship was found for ccf2ap, $t(15) = 2.062, p = .057$, *partial-r* = .470.

Thus, of the four FAUCET measures, the Dominance metric was most predictive of shared mental model accuracy for taskwork and teamwork knowledge. It seems that whether or not team members share knowledge is at least partially determined by the role identities of conversation leaders and followers.

Finally, we predicted that increased interpositional taskwork knowledge and intrateam similarity (SMM Similarity Scores), should correspond to decreases in communication frequency (H6.3). For Experiment 3, knowledge was measured twice. It was measured only once for Experiment 4. In both cases, the knowledge criteria were predicted from total average amount of speech, by all three team members, averaged across all missions. Hypothesis 6.3 was not supported, in that no relationship was found between total speech, and intrateam similarity or interpositional knowledge.

6.8.3 Conclusions

- There is some evidence that taskwork and teamwork SMMs can be tied to LSA measures of team communication content; specifically, lag coherence and transcript density.
- Across both Experiments 3 and 4, Taskwork SMM Similarity Scores were predictive of transcript density. Namely, a higher degree of similarity among the taskwork mental models of team members is indicative of a lower rate of task-specific jargon, but only for co-located teams. One interpretation for this finding is that taskwork similarity may result in less talking and therefore less UAV-specific jargon supporting H6.3
- In Experiment 3, Teamwork SMM Similarity Scores were found to be predictive of the communication measures only for distributed teams. Specifically, teamwork SMMs were positively related to lag coherence whereas teamwork SMMs were negatively related to transcript density, partially supporting H6.3.
- Teams with greater teamwork knowledge also have PLO's that speak as normatively predicted.
- Team taskwork knowledge is better when the team members are independent (in regard to communication flow), with one exception. Teams have better collective taskwork knowledge when PLO reacts to AVO's utterances.
- Teamwork knowledge is highest when AVO's communication is tied to the other teammates, especially DEMPC.

6.9 Task 7: Examine Relation Between Communication and Team Situation Awareness

6.9.1 Method

In order to examine the relationship between FAUCET or content-based communication metrics and TSA (Team Situation Awareness) for Experiments 3 and 4, TSA was measured during each mission via accuracy of either summed or consensus team member responses to situation awareness queries. In order to draw comparisons across experiments, TSA measures from Missions 4 (low workload) and 5 (high workload) were used for each experiment. Summed, or *collective*, accuracy was measured by summing over individually elicited team member responses. For example, each team member is individually queried about the number of targets he or she thinks their team will successfully photograph, and this answer is then compared to the actual outcome, scored and then the score is summed across team members. Consensus, or *holistic*, accuracy was measured by eliciting a response from the team as a whole. For example, the team as a whole is queried about the nature of the next waypoint on their route. Once the team reaches a consensus their answer is compared to the actual state of the environment. The former example is considered a "repeated query" and the latter is an example of a "non-repeated query" in that the former query was given at every mission while the latter was not. Based on a high correlation with team performance, as well as the *development* of

TSA processes over time, only the repeated queries are analyzed here. A significant caveat of interpreting the results of the repeated TSA query however is that we do not fully embrace this a valid measure of TSA because accuracy of response may have a large memorization component involved. That is, the team knows the query is coming and over time they tend to learn the correct response.

Do specific communication patterns lend themselves to better team situation awareness? This question is important in allowing eventual diagnosis of team states (e.g., poor team SA) through communication data. Again, the nature of the environment (co-located vs. distributed) and its impact on the relation between situation awareness and communication is examined, as is changes in situation awareness with experience. How does the nature of the environment (i.e. distributed vs. F2F) affect the development of team situation awareness and how does communication relate to the development of team situation awareness?

We can make some specific predictions about the relationship between team situation awareness and the content of communications as based on LSA content coding. First, teams with higher levels of team situation awareness will follow statements with acknowledgements (H7.1). Although there were no LSA tagging results to directly support H7.1, the previous results that supported H3.5 are relevant here. That is, better performing teams were more likely to follow statements with acknowledgements compared to less effective teams. Given that for Experiments 3 and 4, team situation awareness is highly correlated with team performance, we can infer indirect support for H7.1. It is also predicted that differences in team situation awareness will be reflected by changes in flow patterns (H7.2).

Table 44
Hypotheses Associated with Task 7.

Task/Hypothesis Number	Hypothesis	Supported?
7.1	Teams with higher levels of team situation awareness will follow statements with acknowledgements.	Yes
7.2	Differences in team situation awareness will be reflected by changes in flow patterns.	Yes

6.9.2 Results

6.9.2.1 Transcript density, lag coherence, and team situation awareness. For the query-based measure of TSA described above, a linear model with the independent variables workload (Mission > 4 is high workload in both experiments) and co-located/distributed, and a covariate for the various LSA metrics were tested for relationships with the dependent variable TSA. Eight models were fitted in all (Table 45).

Table 45
Models for Predicting Query-based TSA Using Content Metrics

Collective TSA	Holistic TSA
Transcript density	Transcript density
Experiment 3 (Model 1)	Experiment 3 (Model 5)
Experiment 4 (Model 2)	Experiment 4 (Model 6)
Lag coherence	Lag coherence
Experiment 3 (Model 3)	Experiment 3 (Model 7)
Experiment 4 (Model 4)	Experiment 4 (Model 8)

No significant (i.e., $p < .10$) relationships were found between team communication and TSA (all $p > .27$). Additionally, co-located/distributed was not a significant predictor in any of the models (all $p > .30$). Workload was reliably a significant predictor in all eight models (all $p \leq .003$). Controlling for all of the other factors in the model, higher workload consistently predicted a lower query-based TSA score.

6.9.2.2 Coordinated awareness of situation by teams. Coordinated Awareness of Situation by Teams (CAST) is an interaction-based measure of TSA. Particularly relevant to this project, a CAST manipulation ("roadblock") interfering with team communication was introduced. Specifically, a CAST score was computed for Experiment 3 Mission 6 based on a communication channel glitch manipulation, based on the content and flow of communication in response to the roadblock. For this planned roadblock the communication channel from DEMPC to AVO was cut for five minutes after information about an unplanned target was given to the DEMPC. Figure 11 shows the CAST scoring sheet for two of the teams we observed (we observed 19 total).

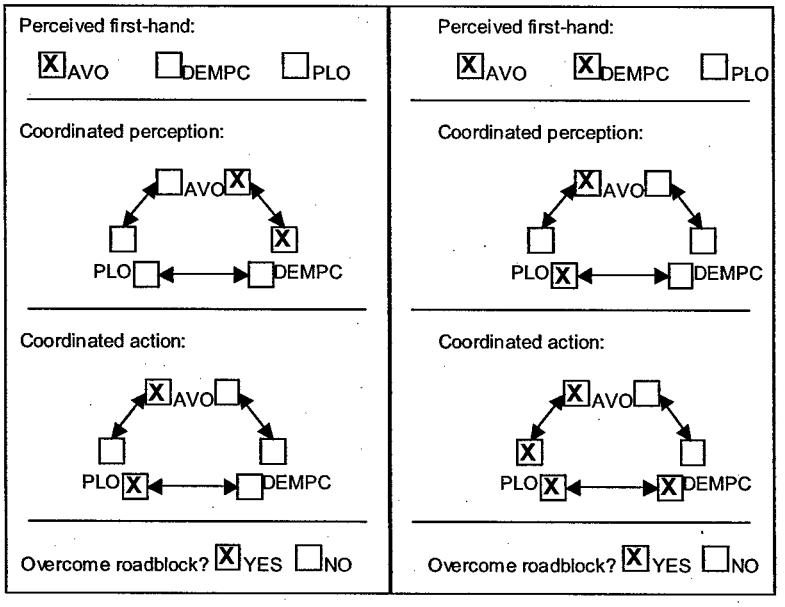


Figure 11. CAST scoring sheet with two sample observations

The score on the left reflects a high TSA score, given that the glitch was addressed ONLY by DEMPC channeling communication to AVO through the PLO in terms of "Coordinated action" (in fact this is the optimal, or "referent," solution). Although the score on the right indicates that this particular team did overcome the roadblock, their TSA score was lower in that AVO channeling communication to DEMPC through the PLO really does not reflect accurate awareness of the current situation. Although CAST measurement implicitly entails communication analysis in this case (i.e., "Coordinated perception" in Figure 11) we ran exploratory pairwise correlations between the ratio of observed to total number of checkboxes for three CAST components, including "Perceived firsthand," "Coordinated perception," and "Coordinated action," and two LSA-based content metrics, lag coherence and transcript density. "Perceived firsthand" was significantly correlated with lag coherence (r (11) = -.51, p = .08). This result may indicate a tendency to have fewer firsthand (as opposed to coordinated/communicated) perceptions across team members when the topic of conversation extends further back in time. The correlation between "Coordinated action" and transcript density was not significant, but was large enough to hint at the possibility of a mild correlation given a larger sample size (r (11) = .21, p = .50). It is important to note that a high rate of team members involved in coordinated action (Figure 11, right panel) does not actually correspond to the changes involved in the unusual situation – in fact this team missed a target because the PLO was too concerned with being a bi-directional conduit between AVO and DEMPC, which really was not necessary. It should be emphasized that these results reflect only one experiment, given that no CAST data were available for Experiment 4.

6.9.2.3 Team situation awareness and communication flow. For both Experiment 3 and Experiment 4, CHUMS inconsistency measures did not predict collective query-based TSA, but did predict holistic query-based TSA. That is, for situation awareness that relies on team consensus, it was important for teams to have a limited number of well-established communication styles, whereas for individual team member SA, this does not appear to be an important factor. Number of CHUMS models (excluding silence as a node in the model, or "No-Null" models) negatively predicted holistic TSA for Experiment 3 (F(1, 13) = 13.827, p = .003, R^2 = .515, B = -0.192), and for Experiment 4 (F(1, 13) = 3.234, p = .095, R^2 = .199, B = -0.202), supporting H7.2.

The ProNet chain length measures of communication stability were also useful in predicting holistic TSA, but were not as consistent as the CHUMS predictors. For both Experiment 3 and Experiment 4, having a lower minimum chain length meant having greater holistic TSA (Experiment 3: F (1, 13) = 4.479, p = .054, R^2 = .256, B = -0.091; Experiment 4: F(1, 14) = 3.439, p = .085, R^2 = .197, B = -0.166). This means that shorter fixed sequences of communication were required to establish a "common ground" in the task, supporting H7.2. These findings were supported by collective TSA in Experiment 3, with a negative relationship for mean, median, and maximum chain length. However, the collective TSA findings were not replicated at Experiment 4.

Taken with the CHUMS results, this implies that TSA – especially holistic TSA – is most

accurate when communication styles are restricted to a few (CHUMS) brief (ProNet) patterns. The implication is that query-based TSA is related to building an established, restricted *flow vocabulary* with which to express current goings-on. These findings are contrary to our performance results, which tend to favor having a few, long patterns, however they are supportive of H7.2. For performance, it might be said that better performing teams are associated with more predictable communication environments.

Both dominance and CRP results revealed relationships between these communication metrics and TSA, but neither relationship was replicated across Experiments 3 and 4. For dominance, TSA was not predictable for Experiment 3. However for Experiment 4 both collective and holistic TSA was fostered by DEMPC relaxing control of the discourse. DEMPC's dominance negatively predicted collective TSA, F (1, 14) = 3.757, p = .073, R^2 = .212, B = -.639. PLO's dominance over DEMPC positively predicted holistic TSA, F(1, 14) = 4.295, p = .057, R^2 = .235, B = 75.685. This suggests that individual understanding of the situation is generally supported by a more "egalitarian" (or at least less DEMPC-driven) discourse. Specifically, team consensus of the situation requires that PLO take a more active role, and that DEMPC respond to this, rather than driving the discourse.

Finally we turn to CRP, which measures communication conformity to a normative model of speech quantity. For Experiment 3 only, AVO and DEMPC deviation from the specified norm predicted better holistic and collective TSA (see Table 46 for inferential statistics). This was not replicated at Experiment 4, however, and it is difficult to interpret these un-replicated findings. The implication is that the normative model of communication quantity is negatively fitted to the team's situation awareness needs. Perhaps this can be taken in conjunction with the negative DEMPC dominance findings from Experiment 4 in which case the implication is that, for TSA purposes, the normative model places too much emphasis on DEMPC (i.e., a single team member's) communication. This also impacts AVO's CRP score as well, since most of the required information passage from DEMPC, goes to AVO.

Table 46
F's and B's for CRP Scores Predicting Query-based TSA in Experiment 3

Collective TSA	
	AVO_CRP: F(1, 11) = 7.581, p = .019, R2 = .408, B = -7.166
	DEMPC_CRP: F(1, 11) = 3.280, p = .097, R2 = .230, B = -3.000
Holistic TSA	
	AVO_CRP: F(1, 11) = 9.432, p = .011, R2 = .462, B = -2.235
	DEMPC_CRP: F(1, 11) = 11.027, p = .007, R2 = .501, B = -1.60

6.9.2.4 FAUCET and CAST TSA. The relationship between the CAST measure of TSA and FAUCET is planned as a future direction, as we are currently collecting more extensive CAST data under a variety of roadblocks. We have come to feel that ultimately CAST measurement of TSA will provide a firmer basis for understanding the relationship

between team communication and TSA. In addition, CAST is very much compatible with FAUCET metrics and we plan to ultimately use the FAUCET metrics to automatically derive the CAST metric.

6.9.3 Conclusions

- There is a tendency to have fewer firsthand (as opposed to coordinated/communicated) perceptions across team members when the topic of conversation extends further back in time.
- For situation awareness that relies on team consensus, it is important for teams to have a limited number of brief, well-established communication styles, supporting H7.2
- Shorter fixed sequences of communication were required to establish a "common ground" in one study, but the finding was not replicated.
- Query-based TSA is related to building an established, restricted *flow vocabulary* with which to express current goings-on. This contradicts performance results, which tend to favor having a few, long patterns, but supports H7.2.
- Individual understanding of the situation is generally supported by a more "egalitarian" (or at least less DEMPC-driven) discourse. This finding was not replicated, however.
- The normative model of communication quantity appears to be negatively fitted to the team's situation awareness needs, but this negative relationship was not replicated. Perhaps the normative model over-emphasizes DEMPC-driven discourse supporting the previous conclusion in regard to Dominance.

7.0 CONCLUSIONS AND IMPLICATIONS

This three year effort generated empirical data that yields theoretical and methodological conclusions, both with implications for understanding and designing for collaboration. This section first outlines the methodological conclusions, followed by theoretical conclusions, limitations, and future directions.

7.1 Measure Validity

In the course of this project, nine different communication analysis measures were applied and evaluated (See Table 47). Four of these measures focused on the analysis of communication content and were based on Latent Semantic Analysis (LSA) and five relied on communication flow - devoid of content. As a reminder, we refer to the collection of flow techniques as FAUCET.

In addition to these measures, there were other methodological innovations in the course of the project. Specifically, in addition to refining our coordination logging and transcription software, we also 1) developed metrics to evaluate tagging agreement for LSA-based automatic tagging and automatic speech recognition case, 2) developed a keyword-based method to provide a "leaner" baseline for LSA, and 3) developed a Web-based LSA interface.

Table 47
Communication Measures Used in Experiments 3 and 4.

MEASURES
New Communication Metrics
Content – LSA-based density
Content- LSA-based performance score
Content – LSA-based automatic tagging
Content- LSA-based lag coherence
Flow - Dominance
Flow – Quantity: CRP
Flow – Sequence: ProNet
Flow – Stability: CHUMS
Flow – Team process surrogate

We hypothesized that our metrics would correspond to team effectiveness (H2.1) and that results would replicate across studies (H2.2). We now have data to support both hypotheses. Our communication analysis metrics generally corresponded to team effectiveness. For instance, our LSA-based performance scores correlated with actual team performance (r =.75) and indeed, this pattern generalized over different semantic spaces and training sets within the same task domain. Similarly the LSA-based tagging procedure was able to consistently predict tags provided by human coders (i.e., content codes) and also generalized over semantic and training spaces. The LSA-based density and coherence functions were also replicated across two of three experiments, though LSA-based density exhibited higher validity than LSA-based lag coherence.

FAUCET methods also fared well on these same criteria. ProNet results indicated that longer detectable sequences (and hence more regular communication patterns) were generally linked to superior performance and process. Further, patterns held across studies. Using CHUMS and ProNet, it was found that communication stability predicted performance. Of the five FAUCET methods, ProNet, Dominance, and CHUMS were more successful in terms of demonstrating validity than CRP and Process Surrogate.

The communication patterns generated by these methods also mapped onto other indices of team performance and cognition (discussed in the next session) further supporting methodological validity.

There are a number of applications of this methodology:

1) *Describing Collaboration.* These measures can be used to describe or summarize through meaningful data reduction, complex behavior like collaboration in a

relatively low dimension measure. This qualitative description alone is useful for research and development in support of collaboration.

2) *Assessing Collaboration.* These methods can go from qualitative patterns to quantitative indices representing patterns that can be used to assess or evaluate team effectiveness. Such evaluative information is invaluable to research and development in support of collaboration and extremely valuable when team-scaled performance measures are unavailable.
3) *Diagnosing Collaboration.* The interpretation of communication patterns can also move beyond assessment to diagnosis, thereby providing a richer explanation for collaboration effectiveness or ineffectiveness. This is accomplished when patterns are tied to collaborative behaviors such as shared mental models or team situation awareness. This information can be used to understand the nature of a team's collaboration strengths or weaknesses for theoretical development and guidance in selecting interventions.
4) *Automation of Collaboration Measures.* However, the value of this approach, relative to its cost may not be immediately apparent to the extent that one can identify other, leaner measures of collaboration that can provide some useful information. Automation of the communication analysis methods would reduce the cost and speed up data analysis time, while providing richer profiles of collaborative behavior. This is the central focus of our current ONR work.
5) *Real-time Communication Analysis.* Not only would collaboration research benefit, but ultimately, the application of this work that seems to be most tantalizing, is the possibility for on-line, real time assessment of team performance and cognition using communication patterns. The use of communication data, an ongoing data stream concomitant with many collaborative tasks, allows for the possibility of automation. That is, the team or group is not interrupted to complete a survey, but rather communication is monitored unobtrusively in real-time. A system that automatically detects problems, such as a breakdown in cohesiveness or lack of team situation awareness, is a pinnacle for team communication researchers and is what uniquely distinguishes this approach from other more obtrusive measure of collaboration.
6) *Disruption of Collaboration.* Finally, there is an application that involves turning the tables on the monitoring and intervention approach. It is possible to conceive applications in which these methods are used to monitor and disrupt enemy collaboration.

Thus, though there are immediate applications of communication analysis techniques, many of the most tantalizing applications are hinging on the automation of the methods, as well as the ability to port methods to different domains. At this point in the project, we have achieved a semi-automated state for most of the methods and this state is in the process of implementation as separate effort. However, in this effort we have also encountered several barriers to ultimate automation and generalizeability to other domains.

Analysis of communication content is the most significant barrier to automation. The first is that unlike the flow-based techniques, the content of discourse (i.e., words,

sentences) is needed and at the moment, it is needed in text format. This creates a need for transcription of the audio-taped discourse. Despite our development of custom transcription software that merged the time-stamped COMLOG (speaker and listener identities) with a window for typing the transcript, this process was a major bottleneck in the effort. The synchronization of the time-stamped COMLOG with the audio discourse was one problem. Without this, the transcription of speaker and listener identity created a large burden on the transcriptionist and without any COMLOG data, these identities (especially listener) are quite difficult to discern. This problem combined with the fact that transcription is a tedious and time consuming process, created an enormous time lag between data collection and post-processed data for content analysis.

Our team has been exploring a solution to this problem in the form of automated speech recognition. Results of this analysis are described under Task 2. In summary, they show that LSA-based performance scores are robust to errors due to speech recognition software. However, this does not solve the problem of speaker and listener identity, information that is needed for some forms of content analysis and which becomes an even greater problem as the number of team members increases from three.

Another barrier to automating content analysis using LSA for use in new domains is the fact that LSA relies on a corpus of domain-specific text. In this case the corpus was a large set of UAV training manuals and other documentation and was later supplemented with our own UAV transcripts. These were used to build the semantic space. Although automation of content analysis could take place within the UAV context of the semantic space, the procedure would be difficult to transfer to other domains requiring a different corpus (e.g., AWACS, emergency operations). On the positive side, results from this effort did demonstrate that the LSA-based results were fairly robust to training sets and semantic spaces derived from different UAV experiments, but this does not address across-domain generality (cf. KeyWI). As a partial solution, keyword indexing software (KeyWi) was written in the course of this effort with a command-line interface. KeyWi is a Java program that takes a corpus input, a keyword input, and file input, in order to produce vector length, distance between utterances, and cosine (i.e., correlation) scores between utterances in a transcript. Although only tested on one experiment, this program is inherently portable relative to our other content methods.

On the other hand, the FAUCET methods lend themselves to complete automation, though there are caveats that should be recognized here. Specifically, for both content and flow techniques the automatic extraction of patterns in the communication data is only a first step in a larger process. The interpretation of these patterns in terms of collaboration effectiveness or more specifically, in terms of team process and cognition, is a different matter. Our interpretation has relied extensively on the existence of other measures of team performance, team situation awareness, and shared mental models. Without these there can be little said about the patterns. In addition, the interpretation is only as good as the criterion measures. This limitation is reflected in our shared mental model and team situation awareness analyses which relied on criterion measures of constructs that were still themselves under empirical scrutiny and development.

Essentially, "good" criterion measures allow better interpretation of the patterns of communication data.

Given good criterion measures and pattern extraction procedures that are automated, there are sill limits to generalizeability beyond the studied domain. A new domain requires not only new criterion measures, but a potentially new interpretation of the patterns, in light of the criterion measures. Ultimately, we envision marrying these pattern extraction procedures and criterion measures with a machine learning procedure that associates, patterns of communication data with the criterion (e.g., effective performance, good team situation awareness) over time. This procedure could be automated, but would require adequate sampling in the data monitoring stage for the machine learning procedure settle on interpretable patterns

One other solution that we envision for eventual automated and efficient pattern extraction and interpretation would involve a stepwise procedure. The first initial screening step would apply the least costly measures first starting with lean measures such as word count and some FAUCET methods. Then, if interesting patterns are observed, some of the more expensive content-based measures could be applied, starting first with keyword methods and then LSA-based methods. The idea would be to use the more powerful, but also more costly methods, only if an interesting pattern is detected using the less costly, shallower methods. Support for this approach was also found in this effort. The keyword method that was developed provided meaningful results comparable to the LSA-based methods and the measure of word count was found to be highly correlated with LSA-vector length, a much more costly measure. However, other results indicated that FAUCET methods predict team performance better than simple, low level communication quantity methods.

In summary, the patterns extracted using our communication analysis methods are predictive of performance in the context of our UAV collaboration. The methods in their current state can inform research on collaboration as described in the following section. Applications in real-time monitoring of collaboration are possible, but will need to address the limitations of transcription (or speech recognition), inadequate criterion measures, and domain dependence.

7.2 Communication and Collaboration

Our communication analysis methods were applied to the team collaboration that happens in the context of a simulated three-person UAV ground control station. Although team communication has been measured before, its richness has seldom been exploited at the level associated with this project. Therefore, in addition to testing our methods, we have had an opportunity to explore in a deep way, the communication behavior of tactical command-and-control teams, the relation of that behavior to team effectiveness and cognition (i.e., shared mental models, team situation awareness), and the effects of distributed work and workload on that communication.

7.2.1 Communication and Team Performance

The results described in the previous section on methodological validity also served as a starting point for understanding the relationship between team communication and performance or effectiveness. We were satisfied by a relationship between communication and performance for the purposes of validity, but under Task 3, explored these relationships more deeply.

In our exploration we found that there are specific patterns of flow and content that are associated with effective performance on our UAV task. Though we have no data to empirically support the *generalizeability* of these results, through careful task analyses and review of the literature we believe that our results likely generalize to similar command-and-control tasks (i.e., very well-structured tasks in which information exchange among a relatively small group predominates).

We have found that effective collaborations in the UAV task can be apparent in what is said. For instance, Markov analyses has shown that teams that tend to state more facts and acknowledge other team members more, tend to perform better, whereas, those that express more uncertainty and need to make more responses to each other tend to perform worse. Also, in terms of content, effective collaborations are associated with higher density (i.e., more "UAV talk") which developmentally starts at a low level, increased dramatically, and tends to return to a moderate level. In addition, as conversations become related over time, we observe increases in team performance.

Examining the flow patterns we see that consistency and regularity in communication flow is a hallmark of effective collaboration and team performance. In addition, we found that increased numbers of communication patterns in order to adapt is also associated with effective collaboration.

These results empirically support what has often been inferred by researchers in the team cognition literature, that expectancies and thus consistencies develop over time and are the hallmark of highly skilled teams.

7.2.2 Communication and Geographic Distribution

In the two studies that contributed to our data sets, the geographic distribution of the three team members was manipulated such that half of the teams were co-located and half were distributed. Note that this manipulation was quite subtle. Namely, even the co-located teams, though they were in the same room sitting in viewing distance, talked over head sets. They were also immersed in their displays and there were rarely face-to-face interactions. The main difference between the co-located and distributed condition was the lack of co-presence, in the latter. That is, those in the distributed condition had no physical awareness of their surrounding team members.

Results of these studies indicated that there were no effects of the manipulation on team performance (Cooke, DeJoode, Pedersen, Gorman, Connor, & Kiekel, 2004).

Interestingly, however, there were effects on team communication. Namely, co-located teams, like effective teams more generally, tended to develop consistent and predictable patterns of communication, also developing their own unique, but consistent team lexicon over time. Co-located teams also exhibited more open (i.e., statement not followed by an acknowledgement or fact) communication channels than distributed teams.

Distributed teams, as opposed to co-located teams, were less consistent in terms of communication patterns. They had more distinct and less stable flow patterns, suggesting that the distributed environment increases the range of behaviors on the part of the team. Similarly, co-located teams demonstrated this type of highly variable behavior (i.e., more communication patterns) in response to the communication break down.

7.2.3 Communication and Workload

Also in the two experiments workload was manipulated by increasing the number of target waypoints that were required to be photographed in a single 40-minute mission. In addition, other parameters of the task environment were manipulated to increase difficulty (e.g., more frequent alarms). Workload, unlike geographic dispersion, had a significant effect on team performance, increasing the time spent per target for higher levels of workload.

Workload also had an effect on communication. First, one might expect that the increased number of targets would also increase the amount of talking in the high workload condition. This is because communication generally needs to occur at, or around, target waypoints. In fact, the opposite occurred. Teams communicated less under high workload (see also Stout, Cannon-Bowers, Salas, & Milanovich, 1999). There were also signs of adaptive behavior occurring under high workload. All teams "behaved" more like distributed teams under these circumstances and adapted by reducing their utterance chain lengths and reversing the dominance pattern typically seen for AVOs and DEMPCs (both are types of flow patterns).

There were also differences in communication content. Under high workload communication was briefer, but more densely packed. Also, the LSA-based approach was able to accurately classify a mission as to whether the team was under high or low workload. This method was found to be robust over variations in the radius and the corpus of team-at-mission transcripts used to make the prediction as well as a variation in which the target names (potentially discriminating targets) were removed from the transcripts. However, the results may also be attributed to the differences between the two types of missions in terms of amount of talking. As reported earlier, the LSA vector length metric used in this scoring approach is correlated with word count.

7.2.4 Communication and Team Cognition

Measures of team knowledge (i.e., shared mental models) and team situation awareness were also taken in the course of data collection for both studies.

Shared mental models of taskwork knowledge appear more directly related to certain aspects of team communication for co-located, compared to distributed teams. Further, this relationship tends to be negative with a higher degree of knowledge similarity related to lower degrees of transcript density. In other words more UAV-speak is used by those teams with little shared knowledge of the task. We can speculate that the increased density may be a compensatory behavior for the low knowledge sharing. Further, shared mental models regarding teamwork appear tied positively to the length of time co-located teams spend on a particular topic (lag coherence). Thus a shared mental model in our UAV task seems to go along with high topic coherence, but low density (UAV-speak). For co-located teams, this pattern of results is less variable and more like that of distributed teams under high workload. The results pertaining to shared mental models and flow patterns were rather weak and sporadic except that dominance seems related to having more knowledge about the task.

Good team situation awareness as measured by the query method tends to be associated with 1) a limited number of brief, well-established communication styles, 2) building an established, restricted *flow vocabulary* with which to express current goings-on, and 3) more egalitarian discourse. It should be noted that this pattern contradicts performance results, which tend to favor having a few, long patterns. This is suggestive of a more general theme that contrasts stable, well established patterns and high team performance with flexible, varied patterns and the ability to adapt communication patterns to novel situations. Teams with high SA, teams that are distributed, teams faced with a communication break down, and teams that are under high workload seem to accommodate the latter pattern.

In general, the results for team cognition may be speculative to the extent that the criterion measures are limited. Previous work has indicated that shared mental models do not explain much of the variance in team performance in the UAV task, although admittedly it is critical to achieve a certain baseline understanding of the task and team. Further, the query-based situation awareness measures do not seem to directly reflect the truly collaborative nature of team situation awareness as much as our new CAST (Coordinated Awareness of the Situation in Teams) score does.

7.3 Naval Relevance

The methods developed and tested here for the analysis of communication will provide tools that will enhance our understanding of the collaborative process and provide a means of identifying the strengths and weaknesses in such processes. At the same time, the further automation and generalization of the computational tools is an essential step toward a system that monitors communication in real-time in order to identify anomalous patterns and intervene to prevent team ineffectiveness or error. Such a system would work in the background of a larger command-and-control system, or shipboard command information system. Other applications include the monitoring of enemy communications to detect patterns and disrupt collaboration in a language-independent way. More immediately, these techniques can be useful for the evaluation of techniques,

systems, or decision aids that are purported to enhance collaboration, for example by contrasting enhanced communication vs. un-enhanced communication.

7.4 Future Directions

Our future work will move toward automation and generality by : 1) implementing some of the most successful FAUCET methods (i.e., ProNet and Dominance) in a software tool, 2) applying the methods to communication analysis in domains that are characterized by longer term planning and decision making, and 3) exploring techniques for automated interpretation of communication patterns. We have also noticed that communication analysis has gained momentum in the years since we began this project with other investigators proposing parallel approaches (e.g., Diedrich, Freeman, Entin, & MacMillan, 2005; Swoboda, Kilduff, & Katz, 2005).

Our findings with regard to communication and collaboration have directed our attention to the importance of pattern stability (and instability) in the collaborative process. Communication patterns are either rigid and stable or variable, flexible, and adaptive. Although high levels of performance can be achieved with the rigid and stable collaborative process, most dynamic situations seem to call for more adaptive patterns. We have also been impressed by teams' propensity to self organize. That is, most patterns, whether stable or variable, are not directly trained, but emerge as a result of team interactions over time. These findings are extremely relevant to team training and subsequent deployment. These and other similar results have led us to think about team cognition as an ecological phenomenon. We see team cognition as emerging from the adaptive interactions of teammates, rather than a collection of cognitive entities (i.e., the individually-trained personnel). This new perspective has led to new theorizing and the development of new measures of team situation awareness (i.e., CAST), as well as measures of team coordination (Cooke & Gorman, in press). We see a strong mapping between communication flow and self-organized aspects of team coordination, and envision that the automation of the flow techniques should ultimately facilitate the automated measurement of self-organized team coordination. Thus, the focus on communication, a team process, has generated a new way of thinking about team cognition and new methods for measuring it. In the long run, the propensity to self-organize (given any team) may weigh heavy on team performance, perhaps even more than individually-based training.

8.0 REFERENCES

Achille, L. B., Schulze, K. G., & Schmidt-Nielsen, A. (1995). An analysis of communicatión and the use of military terms in Navy team training, *Military Psychology, 7*, 95-107.

Beebe, S. A., and Masterson, J. T. (1997). *Communicating in Small Groups*. (5th Ed). New York, NY: Longman.

Berry, M. W. (1992). Large scale singular value computations. *International Journal of Supercomputer Applications, 6,* 13-49.

Bowers, C. A., Braun, C. C., & Kline, P. B. (1994). Communication and team situational awareness. In R.D. Gilson, D. J. Garland, & J. M. Koonce (Eds.), *Situational Awareness in Complex Systems* (pp. 305-311). Daytona Beach, FL: Embry Riddle Aeronautical University Press.

Bowers, C. A., Jentsch, F., Salas, E., & Braun, C. C. (1998). Analyzing communication sequences for team training needs assessment. *Human Factors, 40,* 672-679.

Bowers, C. A., Urban, J. M., & Morgan, B. B., Jr. (1992). *The study of crew coordination and performance in hierarchical team decision making* (Rep. No. TR-92-01). Orlando, FL: University of Central Florida Team Performance Laboratory.

Carletta, J. (1996). Assessing agreement on classification tasks: the kappa statistic. *Computational Linguistics, 2,* 249-254.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20, 34-46.

Contractor, N. S., & Grant, S. J. (1996). . The emergence of shared interpretations in organizations: a self-organizing systems perspective. In J. H. Watt and A.C. VanLear (Eds.), *Dynamic Patterns in Communication Processes*, pp. 215-230. Thousand Oaks, CA: Sage Publications.

Cooke, N. J., DeJoode, J. A., Pedersen, H. K., Gorman, J. C., Connor, O. O., & Kiekel, P. A. (2004). The Role of Individual and Team Cognition in Uninhabited Air Vehicle Command-and-Control. Technical Report for AFOSR Grant Nos. F49620-01-1-0261and F49620-03-1-0024.

Cooke, N. J. & Gillan, D. J. (1999). Representing user behavior in human-computer interaction. In A. Kent & J. G. Williams (Eds.), *Encyclopedia of Computer Science and Technology,* pp. 283-308. New York: Marcel Dekker, Inc. Also to be reprinted in the *Encyclopedia of Library and Information Science.*

Cooke, N. J., & Gorman, J. C. (in press). Assessment of team cognition. In P. Karwowski (Ed.), *2nd EDITION- International Encyclopedia of Ergonomics and Human Factors*. Taylor & Francis Ltd.

Cooke, N. J., Neville, K. J., & Rowe, A. L. (1996) Procedural network representations of sequential data. *Human-Computer Interaction, 11,* 29-68.

Cooke, N. J., Rivera, K., Shope, S.M., & Caukwell, S. (1999). A synthetic task environment for team cognition research. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting,* 303-307.

Cooke, N. J., Salas, E., Cannon-Bowers, J. A., & Stout, R. (2000). Measuring team knowledge. *Human Factors, 42,* 151-173.

Cooke, N. J., & Shope, S. M. (2002). The CERTT-UAV Task: A Synthetic Task Environment to Facilitate Team Research. *Proceedings of the Advanced Simulation*

Technologies Conference: Military, Government, and Aerospace Simulation Symposium, pp. 25-30. San Diego, CA: The Society for Modeling and Simulation International.

Cooke, N. J., Shope, S.M., & Rivera, K. (2000). Control of an uninhabited air vehicle: A synthetic task environment for teams. *Proceedings of the Human Factors and Ergonomics Society 44th Annual Meeting*, 389.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of he American Society For Information Science, 41,* 391-407.

Diedrich, F. J., Freeman, J., Entin, E. E. & MacMillan, J. (2005). Modeling, measuring, and improving cognition at the team level. Paper presented at the First Augmented Cognition International Conference, July 22-28, Las Vegas, NV.

Di Eugenio. B. (2000). On the usage of Kappa to evaluate agreement on coding tasks. *LREC2000, the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 2000.

Donaghy, W. C. (1989). Nonverbal communication measurement, in P. Emmert and L. L. Barker (Eds.), *Measurement of Communication Behavior*, pp. 296-332. White Plains, NY: Longman, Inc.

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., & Nikolic, D. & Manning, C. A. (1998). Situation awareness as a predictor of performance in en route air traffic controllers. *Air Traffic Control Quarterly.*

Emmert, V. J. (1989). Interaction analysis in P. Emmert and L. L. Barker (Eds.), *Measurement of Communication Behavior*, pp. 218-248. White Plains, NY: Longman, Inc.

Emmert, P., & Barker, L. L. (1989). *Measurement of Communication Behavior.* White Plains, NY: Longman, Inc.

Foltz, P. W. (1996) Latent Semantic Analysis for text-based research. *Behavior Research Methods, Instruments and Computers. 28*, 197-202.

Foltz, P. W., Kintsch, W. & Landauer, T. K. (1998). The measurement of textual coherence with Latent Semantic Analysis. *Discourse Processes, 25,* 285-307.

Foltz, P. W., Laham, D., & Landauer, T. K. (1999). The Intelligent Essay Assessor: Applications to Educational Technology. *Interactive Multimedia Education Journal of Computer Enhanced Learning. 1,* (2).

Foushee, H. C., & Manos, K. (1981). Information transfer within the cockpit: Problems in intracockpit communications. In C. E. Billings & E. S. Cheaney (Eds.), Information transfer problems in the aviation system (Report No. NASA TP-1875). Moffett Field, CA: NASA-Ames Research Center.

Gillan, D.J., and Cooke, N. J. (in press). Using Pathfinder networks to analyze procedural knowledge in interactions with advanced technology. In E. Salas (Ed.), Human/*Technology Interaction in Complex Systems.* Greenwich, CT: JAI Press Inc., Vol. 10.

Grove, W. M., Andreasen, N. C., McDonald-Scott, P., Keller, M. B., & Shapiro, R. W. (1981). Reliability studies in psychiatric diagnosis: Theory and practice. *Archives of General Psychiatry, 38*, 408-413.

Gugerty, L., DeBoom, D., Walker, R., & Burns, J. (1999). Developing a simulated uninhabited aerial vehicle (UAV) task based on cognitive task analysis: Task

analysis results and preliminary simulator data. *Proceedings of the Human Factors and Ergonomics Society 43rd Annual Meeting* (pp. 86-90). Santa Monica, CA: Human Factors and Ergonomics Society.

Harman, D. (1986). An experimental study of the factors important in document ranking. In *Association for Computing Machinery Conference on Research and Development in Information Retrieval,* . Association for Computing Machinery.

Holmes, M. (1997). Optimal matching analysis of negotiation phase sequences in simulated and authentic hostage negotiations. *Communication Reports, 10,* 1-8.

Howell, W. C., & Cooke, N. J. (1989). Training the human information processor: A look at cognitive models. In I. Goldstein (Ed.), Training *and Development in Work Organizations: Frontier Series of Industrial and Organizational Psychology, Volume 3, New* York: Jossey Bass, 121-182.

Hutchins, E. (1995). *Cognition in the Wild.* Cambridge, MA: MIT Press.

Jentsch, F. G., Sellin-Wolters, S., Bowers, C.A., & Salas, E. (1995). Crew coordination behaviors as predictors of problem detection and decision making times. In *Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting* (pp. 1350-1353). Santa Monica, CA: Human Factors and Ergonomics Society.

Kiekel, P. A., Cooke, N. J., Foltz, P. W., & Shope, S. M. (2001). Automating measurement of team cognition through analysis of communication data. In M. J. Smith, G. Salvendy, D. Harris, and R. J. Koubek (Eds.), *Usability Evaluation and Interface Design,* pp. 1382-1386, Mahwah, NJ: Lawrence Erlbaum Associates.

Kleinman, D. L., & Serfaty, D. (1989). Team performance assessment in distributed decision making. *Proceedings of the Symposium on Interactive Networked Simulation for Training* (pp. 22-27). Orlando, FL: University of Central Florida.

Klimoski, R., & Mohammed, S. (1994). Team mental model: Construct or metaphor? *Journal of Management, 20,* 403-437.

Krippendorff, K. (1980). Content Analysis: an Introduction to its Methodology. Beverly Hills: Sage Publications.

Landauer, T. K. & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104,* 11-140

Landauer, T. K, Foltz, P. W. & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25,* 259-284.

Merritt, A. C., Helmreich, R. L. (1996). Human factors on the flight deck: The influence of national culture. *Journal of Cross-Cultural Psychology, 27,* 5-24.

Mosier, K. L., & Chidester, T. R. (1991). Situation assessment and situation awareness in a team setting. In Y. Quéinnec & F. Daniellou (Eds.), *Designing for everyone: Proceedings of the 11th Congress of the International Ergonomics Association* (pp. 798-800). London: Taylor & Francis.

Norman, D. A. (1986). Cognitive engineering. In D. A. Norman and S. W. Draper (Eds.), *User centered system design* (pp. 31-61). Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Orasanu, J. (1990). *Shared mental models and crew performance* (Report No. CSLTR-46). Princeton, NJ: Princeton University.

Oser, R. L., Prince, C., Morgan, B. B., Jr., & Simpson, S. (1991). *An analysis of aircrew communication patterns and content* (NTSC Tech. Rep. No. 90-009). Orlando, FL: Naval Training Systems Center.

Poole, M. S., Holmes, M., Watson, R., & DeSanctis, G. (1993). Group decision support systems and group communication: a comparison of decision making in computer-supported and nonsupported groups. *Communication Research, 20*, 176-213.

Rehder, B., Schreiner, M. E., Wolfe, B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using Latent Semantic Analysis to assess knowledge: Some technical considerations. *Discourse Processes. 25*(2&3).

Rietveld, T. & van Hout, R. (1993). *Statistical Techniques for the Study of Language and Language Behavior.* Mouton de Gruyter.

Sanderson, P. M. & Fisher, C. (1994). Exploratory sequential data analysis: Foundations. *Human Computer Interaction, 9*, 251-317.

Salas, E., Bowers, C. A., & Cannon-Bowers, J. A. (1995). Military team research: 10 years of progress. *Military Psychology, 7*, 55-75.

Salas, E. Cannon-Bowers, J.A., Church-Payne, S., & Smith-Jentsch, K. A. (1998). Teams and teamwork in the military. In C. Cronin (Ed.), *Military Psychology: An Introduction* (pp. 71-87). Needham Heights, MA: Simon & Schuster.

Schmidt-Nielsen, A. Marsh, E. Tardelli,, J. Gatewood, P., Kreamer, E., Tremain, T., Cieri, C. , Strassel,S. Martey,N., Graff, D., Tofan, C. (2001). Speech in Noisy Environments (SPINE2) Part 1 Audio . Linguistics Data Consortium catalog: LDC2001S04.

Schvaneveldt, R. W. (1990). *Pathfinder associative networks: Studies in knowledge organization.* Norwood, NJ: Ablex.

Smith, J. B. (1994). *Collective Intelligence in Computer-Based Collaboration.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Stout, R. J. (1995). Planning effects on communication strategies: A shared mental model perspective. *Proceedings of the Human Factors Society 39th Annual Meeting* (pp. 1278-1282).

Stout, R.J., Cannon-Bowers, J.A., Salas, E., and Milanovich, D.M. (1999). Planning, shared mental models, and coordinated performance: An empirical link is established. *Human Factors, 41*, 61-71.

Sundstrom, E., DeMeuse, K. P., & Futrell, D. (1990). Work teams: Applications and effectiveness. *American Psychologist, 45*, 120-133.

Swoboda, J.C., Kilduff, P. W. & Katz, J. P. (2005). A platoon level model of communication flow and the effects on soldier performance. *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting,* Orlando, FL, pp. 1210-1214.

Thornton, R. C. (1992). *The effects of automation and task difficulty on crew coordination, workload, and performance.* Unpublished doctoral dissertation, Old Dominion University, Norfolk, VA.

Walther, J. B. (1996). Computer-mediated communication: Impersonal, interpersonal, and hyperpersonal interaction. *Communication Research, 23,* 3-43.

Watt, J. H. & VanLear, A. C. (1996). *Dynamic Patterns in Communication Processes.* Thousand Oaks, CA: Sage Publications.

Wegner, D. M. (1995). A computer network model of human transactive memory. *Social Cognition, 13,* 319-339.

Wolfe, M., B. Schreiner, M. E., Rehder, B., Laham, D., Foltz, P. W., Kintsch, W. & Landauer, T. K (1998). Learning from text: Matching readers and texts by Latent Semantic Analysis. *Discourse Processes, 25*, 309-336.

Woods, D. D. & Roth, E. M. (1988). *Cognitive systems engineering.* In M. Helander (Ed.), Handbook of Human-Computer Interaction (3-43). Amsterdam: Elsevier Science Publishers BV.

9.0 BIBLIOGRAPHY

Publications Resulting From ONR-Funded Effort
(in chronological order)

Cooke, N. J., Salas, E., Kiekel, P. A., & Bell, B. (2004). Advances in measuring team cognition. In E. Salas and S. M. Fiore (Eds.), Team Cognition: Understanding the Factors that Drive Process and Performance, pp. 83-106, Washington, DC: American Psychological Association.

Gorman, J.C., Cooke, N.C., & Kiekel, P.A. (2004). Dynamical Perspectives on Team Cognition. Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting.

Kiekel, P. A. & Cooke, N. J. (2004). Human factors aspects of team cognition. In R. W. Proctor and K. L. Vu (Eds.), The Handbook of Human Factors in Web Design, pp. 90-103, Mahwah, NJ: Lawrence Erlbaum Associates.

Kiekel, P.A., Gorman, J.C., & Cooke, N.C. (2004). Measuring Speech Flow of Co-located and Distributed Command and Control Teams During a Communication Channel Glitch. Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting.

Kiekel, P.A. (2004). Developing automatic measures of team cognition using communication data. Ph.D. Thesis, New Mexico State University.

Cooke, N. J. (2005). Measuring Team Knowledge. Handbook on Human Factors and Ergonomics Methods, pp. 49-1-49-6. Boca Raton, FL: CLC Press, LLC.

Cooke, N. J. (2005). Augmented Team Cognition. Paper presented and session chaired at Augmented Cognition International, HCII, July 22-27, Las Vegas, NV.

Foltz, P. F. (2005). Tools for Enhancing Team Performance through Automated Modeling of the Content of Team Discourse. Paper presented at Augmented Cognition International, HCII, July 22-27, Las Vegas, NV.

Foltz, P. W. (2005) Automated Content Processing of Spoken and Written Discourse: Text Coherence, Essays and Team Analyses. Document Design., 13,1, pp. 5-13.

Gorman, J. C. (2005). The Concept of Long Memory in Assessing the Global Effects Augmented Team Cognition. Poster presented at Augmented Cognition International, HCII, July 22-27, Las Vegas, NV.

Gorman, J. C., Cooke, N. J., Pedersen, H. K., Connor, O. O., & DeJoode, J. A. (2005) Coordinated Awareness of Situation by Teams (CAST): Measuring Team Situation Awareness of a Communication Glitch. Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting, 274-277.

Kiekel, P. A. (2005). FAUCET: Using Communication Flow Analysis to Diagnose Team Cognition. Paper presented and session chaired at Augmented Cognition International, HCII, July 22-27, Las Vegas, NV.

Kiekel. P. A., & Winner, J. (2005). Lag Sequential Analysis Using PRONET: Effects of Series Length and Noise. Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting, 2115-2118.

Gluck, K. A., Ball, J. T., Gunzelmann, G., Krusmark, M.A., Lyon, D. R., & Cooke, N. J. (2006). A Prospective Look at Synthetic Teammate for UAV Applications. Invited talk for AIAA "Infotech@Aerospace" conference on Cognitive Modelling.

Cooke, N. J., & Gorman, J. C. (in press). Assessment of team cognition. In W. Karwowski (Ed.), 2nd EDITION- International Encyclopedia of Ergonomics and Human Factors. Taylor & Francis Ltd.

Cooke, N. J., Gorman, J. C., & Winner, J. L. (submitted). Team cognition. In F. Durso, R. Nickerson, S. Dumais, S. Lewandowsky, and T. Perfect, *Handbook of Applied Cognition, 2nd Edition* Wiley.

Gorman, J.C.., Cooke, N. J., & Winner, J.L. (submitted). Measuring team situation awareness in decentralized command and control systems. Submitted to *Ergonomics.*

10.0 TRANSITIONS

Under technology transfer we include the application of methods developed primarily for this project to other, outside projects as well as the development of new technologies based on the research under this grant.

- Cooke and students have been working with a small business in Tempe called Crawdad. We have shared some of our communication data with Kevin Dooley and Steve Corman of Crawdad and ASU, who have been applying centering resonance analysis to these data.

- Peter Foltz has discussed the results of these studies with personnel in the Navy (Katie Ricci, NAVAIR, Ray Perez, ONR), as well as with personnel from DARPA (Ralph Chatham), AFRL (Winston Bennett, HEAA), and ARL (Linda Pierce, Mike Strub). Some of the methods developed as part of the research will be tested within an Army peace-keeping context funded through ARL.

- Peter Foltz is also working to transfer some of the LSA-based communication analysis measures through Pearson Knowledge Technologies. They are currently working on contracts with AFRL, ONR and DARPA, which can benefit from such technologies.

- Under development of new technology, we have a new user-friendly web-based LSA tool, which can be used to assess team communication against a UAV semantic space ("Latent Semantic Analysis in Action"; http://bluff.nmsu.edu/~ahmed/).

- We have also under this effort refined the transcription software that merges the comlog data that generates speaker, listener and time stamp with a window for transcription.

- The CERTT Lab is working in conjunction with Kevin Gluck's Palm Lab at the Air Force Research Lab. The CERTT Lab is providing communication data that can serve as a target for a natural language processing front end of an intelligent agent who will serve as the simulation AVO.

- Communication flow methods are being incorporated into a communication analysis tool in a joint project with Aptima and Kathleen Carley (CMU) for the Navy. Some of the CERTT Lab's transcribed data has also been transitioned to this group for testing analytic methods.

- The CERTT (Cognitive Engineering Research on Team Tasks) Laboratory has relocated to a new facility (Cognitive Engineering Research Institute) and continues to host hundreds of visitors each year for demonstrations and tours.

- Dr. Cooke has discussed this work in invited talks at Air Force Research Laboratory, Georgia Tech, University of North Dakota, and Texas Tech, as well as numerous conferences and small meetings including:
    - A meeting sponsored by Human Factors and Ergonomics Society and the Federation of Social and Behavioral Science on Human Factors of Homeland Security
    - A National Academies Workshop on Scalable Interfaces for Air and Ground Military Robots.
    - A team workshop sponsored by University of Central Florida and the Army Research Institute
    - Two CERI-sponsored Human Factors of UAVs workshops (May 2005 and May 2005)

- Dr. Cooke also has a statement about coordination in hurricane Katrina on the APA web site (http://www.apa.org/ppo/issues/katrinaresearch.html).

- Dr. Cooke is also on two NRC National Academies of Science committees in which team coordination and the communication metrics have been discussed.

Related Projects

- Air Force Office of Scientific Research and Air Force Research Laboratory grant to Cooke; Acquisition and Retention of Coordination in Command-and-Control. This is an integrated empirical and modeling effort to understand, model, and measure team coordination as it evolves with skill acquisition and periods of disuse.

- STTR N04-T026 AP-P-523 (subcontract to Cooke at CERI from Aptima) IMAGES: Instrument for the Measurement and Advancement of Group Environmental SA. Involves applying communication flow techniques to larger communication analysis tool.

- Army Research Laboratory Advanced Decision Architecture Collaborative Technology Alliance (subcontract from MicroAnalysis and Design) to Foltz and others. Research contract to investigate culture, communications and cognition of teams doing intelligence decision-making tasks.

- Air Force Office of Scientific Research to Foltz. Automatic Communication Analysis System using Latent Semantic Analysis. Research grant to study communication analyses of Air Force communications from distributed mission training environments.